Retrospective Approximation
for Smooth Stochastic Optimization

David Newton

Purdue University, Department of Statistics, newton34@purdue.edu,

Raghu Bollapragada

The University of Texas at Austin, Department of Mechanical Engineering, Raghu.bollapragada@utexas.edu,

Raghu Pasupathy

Purdue University, Department of Statistics, pasupath@purdue.edu,

Nung Kwan Yip

Purdue University, Department of Mathematics, yipn@purdue.edu,

Stochastic Gradient (SG) is the defacto iterative technique to solve stochastic optimization (SO) problems with a smooth (non-convex) objective f and a stochastic first-order oracle. SG's attractiveness is due in part to its simplicity of executing a *single step* along the negative subsampled gradient direction to update the incumbent iterate. In this paper, we question SG's choice of executing a *single step* as opposed to *multiple steps* between subsample updates. Our investigation leads naturally to generalizing SG into Retrospective Approximation (RA) where, during each iteration, a "deterministic solver" executes possibly *multiple steps* on a subsampled deterministic problem and stops when further solving is deemed unnecessary from the standpoint of statistical efficiency. RA thus rigorizes what is appealing for implementation — during each iteration, "plug in" a solver, e.g., L-BFGS line search or Newton-CG, *as is*, and solve only to the extent necessary. We develop a complete theory using *relative error* of the observed gradients as the principal object, demonstrating that almost sure and L_1 consistency of RA are preserved under especially weak conditions when sample sizes are increased at appropriate rates. We also characterize the iteration and oracle complexity (for linear and sub-linear solvers) of RA, and identify a practical termination criterion leading to optimal complexity rates. To subsume non-convex f, we present a certain "random central limit theorem" that incorporates the effect of curvature across all first-order critical points, demonstrating that the asymptotic behavior is described by a certain mixture of normals. The message from our numerical experiments is that the ability of RA to incorporate existing second-order deterministic solvers in a strategic manner might be important from the standpoint of dispensing with hyper-parameter tuning.

Key words: Sample Selection, Machine Learning
MSC2000 subject classification:
OR/MS subject classification: Primary: ; secondary:
History:

1. INTRODUCTION We propose iterative algorithms to solve the unconstrained smooth stochastic optimization problem

$$\underset{x \in \mathbb{R}^d}{\text{minimize}} \quad f(x) := \mathbb{E}[F(x,Y)] = \int F(x,y)\,P(dy), \qquad (Q)$$

where Y is a random variable having distribution P on a measurable space $(\mathscr{Y}, \mathscr{A})$, and $F(\cdot, \cdot)$: $\mathbb{R}^d \times \mathscr{Y} \to \mathbb{R}$ is a "random function". Our standing assumption is that the function F is smooth while possibly non-convex, and that we have access to a stochastic first-order blackbox oracle which allows "observing" the function F and its gradient ∇F at any $x \in \mathbb{R}^d$ and $Y = y \in \mathscr{Y}$. (We will make precise the notion of a stochastic first-order oracle and the underlying sigma algebra within the algorithm in Sections 2 and 3 respectively.) Optimization problems of this form arise in a variety of applications arising within simulation optimization [8, 23, 36, 53, 54], machine learning [13], reinforcement learning [6, 43, 62], data fitting [22] and many others.

1.1. Stochastic Gradient and Sample Average Approximation Stochastic Gradient (SG) [13, 34] (originally called Stochastic Approximation in the seminal work of Robbins and Monro [58]) and Sample Average Approximation (SAA) [35, 70] are the two main algorithmic frameworks used for solving (Q). (See [13, 34, 35, 70] and references therein for extensive commentary on the respective strengths and limitations of SG and SAA.) In its most basic form, SG starts with an initial guess, and repeatedly updates the parameters x by executing a gradient-descent like recursion using a sampled approximation of $f(x)$ at each iteration. Over the last decade or so, a number of variations of SG have been developed and demonstrated to be optimal. For example, mirror descent stochastic approximation [48] and accelerated stochastic approximation [40] exhibit optimal complexity [40] on stochastic convex optimization problems. Similarly, the $O(\epsilon^{-4})$ complexity exhibited by randomized stochastic gradient [24] on non-convex smooth stochastic optimization problems is optimal since it has been recently demonstrated [3] that $O(\epsilon^{-4})$ forms a lower bound complexity for smooth non-convex stochastic optimization using a first-order oracle.

Despite such guarantees of optimality, work on SG has continued over the decades mostly in attempts to make SG work well in practice without the need for tuning hyperparameters. For example, the most recent wave of progress in SG has resulted in the incorporation of such features as momentum [37], adaptive stepsizes [41, 79], variance reduction techniques (SAG [63] SAGA [19], SVRG [33]), stochastic higher-order methods to incorporate curvature information via quasi-Newton update procedures [5, 11, 46, 64] or via subsampled Hessian approximations [10, 54, 59], and adaptive sampling ideas [9, 15, 16, 27] aimed at gradually improving the quality of the approximation of f as the iterates evolve. SG approaches, by their nature, tend to produce iterates that have low bias but high variance, low computational cost per iteration but inferior iteration complexity, and the need for tuning of hyper-parameters [18].

REMARK 1.   We have refrained from using the popular misnomer "Stochastic Gradient Descent" in favor of the more appropriate "Stochastic Gradient" since SG is not a descent algorithm in that no guarantees of progress are available during each step.

A broad competing framework to SG is sample-average approximation is SAA. In contrast to SG, advances in SAA, especially its specialized optimization-oriented development, has taken place only since the 1990s [28, 60, 66, 67, 68, 69, 70, 71] due largely to the seminal work of A. Shapiro. (We note that the roots of SAA can be traced back to M-estimation [30, 61, 72] from classical statistics.) The important conceptual idea within SAA is the notion of a *sample-path approximation* f_m constructed using an implicitly generated fixed sample $Y_1, Y_2, \ldots, Y_m$ in an attempt to approximate the objective f, and on which a chosen deterministic optimization algorithm is executed to high accuracy to obtain an estimator X_m^* of a solution to (Q). Traditionally, investigations in SAA focus on inferential questions about relevant SAA estimators, e.g., $f_m(X_m^*)$, X_m^*, and $f(X_m^*)$, under different structural conditions of the objective f and its gradient ∇f, and the nature of the stochastic oracle at hand.

Unsurprisingly, SAA solution estimators tend to be "biased towards" the chosen sample [42], tend to have high per-iteration computational cost due to using a high sample size for approximating f, but also tend to enjoy superior iteration complexity. A crucial practical advantage that SAA enjoys over SG is that well-established deterministic optimization algorithms such as L-BFGS [51] and Newton-CG [51] can simply be "plugged in" as a solver in the service of optimizing f_m.

1.2. Retrospective Approximation (RA)   The framework we present here, called "Retrospective Approximation (RA)" is designed to combine advantageous elements within SG and SAA, much like some ideas that have appeared in the literature over the last two decades [21, 29, 31, 32, 65]. RA has also appeared before, in the context of solving two-stage stochastic programming problems [56] and stochastic root-finding problems [17, 52]. RA is organized into outer and inner iterations, where during each outer iteration, the incumbent iterate is updated by executing a deterministic optimization algorithm on a sample-path approximation of the objective, for a specified number of inner iterations.

Three aspects of RA are salient. First, unlike in SA where a *single* update of the iterate is performed using each sample generated for function and gradient approximation, RA allows *multiple* updates of the iterate before considering a new sample for function and gradient approximation. Through such design, and assuming the use of common random numbers (see Section 3), RA enjoys SAA's facility to "plug-in" existing quasi-Newton deterministic optimization algorithms such as L-BFGS and Newton-CG for performing the iterate updates. Such ability to directly use existing solvers avoids complex redesign of optimization algorithmic components such as steplength selection and Hessian estimation, and the necessity of tuning hyper-parameters, a procedure that is often necessary in SG [18].

Second, RA deviates from traditional SAA in that it explicitly considers the question of how many updates to perform before considering a new sample to approximate the function and its gradient. Comparing the statistical benefit derived from one additional iterate update versus a sample update is essentially a formal treatment of the "estimation error versus optimization error" question discussed in [13], but within a sequential adaptive SAA context. RA answers this question by incorporating an adaptive stopping tolerance which dictates when it might be advantageous to stop iterate updates in favor of a sample update.

A third aspect that works in favor of RA is the potential reduction in the number of times a new approximation of f (or its gradient) is constructed, savings that might be significant in machine learning settings where such approximations might incur nontrivial computational cost.

REMARK 2.   The recent work in [31] requires special mention due to its similar point of view with RA in attempting to incorporate a quasi-Newton scheme alongside increasing sample sizes in the service of solving a stochastic optimization algorithm. However, RA differs in two respects that we think are crucial for both theoretical and implementation standpoints: (i) as we clarify in Section 3, RA *fixes* the sample between outer iterations, allowing a deterministic quasi-Newton solver to retain and exploit structure in the resulting deterministic optimization problem. In contrast, the quasi-Newton scheme in [31] relies on stochastic updating of the Hessian matrix obtained using *differing* samples over iterations; and (ii) the deterministic solver being used in the outer iteration of RA is stopped using an adaptive error tolerance rule that judges whether it might be statistically efficient to continue optimizing the current sample-path problem.

1.3. Summary of Contributions   Considering the vast existing literature and the need to succinctly delineate RA's position, we summarize here the key aspects of the paper.

(a) We propose Retrospective Approximation (RA) as a framework to solve smooth non-convex stochastic optimization problems using a stochastic oracle. RA uses a chosen quasi-Newton solver, e.g., LBFGS or Newton-CG, to solve a sequence of sampled deterministic optimization problems to a specified adaptive tolerance chosen to trade-off estimation error and optimization error. RA can also be viewed differently, as a generalized form of SG that incorporates the potential for multiple solution updates between sample updates.

(b) We demonstrate that almost sure and L_1 consistency of iterates generated by RA are preserved under especially weak conditions. We characterize the iteration and oracle complexity when linear and sub-linear solvers are used within RA, and identify a practical solver termination criterion leading to optimal complexity rates.

(c) Our general premise is that controlling the *relative error* of the observed gradients, as opposed to their *absolute error*, is sufficient to ensure desirable behavior of algorithms for stochastic optimization. Accordingly, our entire mathematical treatment assumes a uniform law on the *relative error* of the observed gradients in contrast with the more typical assumption of a uniform law on the *absolute error* [4, 12, 44, 45, 70, 77]. As we demonstrate in Section 6, such treatment results in increased generality, allowing to subsume important problem classes otherwise excluded by a uniform law on the absolute error.

(d) The objective f is smooth non-convex and to reflect the effect of curvature at all first-order critical points of f, we establish a weak convergence result (a certain "random CLT") on the sequence of RA iterates, and the sequence of true gradients observed at RA's iterates.

(e) Extensive numerical testing suggests that RA, due to its facility of incorporating established deterministic solvers without modification, seems to be able to dispense with hyper-parameter tuning. In particular, our numerical results demonstrate that higher-order methods when implemented within the RA framework are competitive to well-tuned SG methods. (All numerical tests in Section 9 were performed using RA embedded with an L-BFGS line search solver.)

(f) Our use of relative error as the principal mathematical object combined with a reliance on some powerful recent results from empirical process theory [26, 75, 76] might present a standard route to generalizing various results on empirical risk minimization that have appeared in the machine learning literature.

1.4. Paper Organization   The paper is organized into nine sections. Section 2 provides background, notation, and common assumptions used in the paper and Section 3 describes the RA framework. Sections 4–8 develop the main ideas of the paper: Section 4 presents almost sure and L_1 consistency results for RA, Section 5 presents iteration complexity results, Section 6 presents four problem classes subsumed by RA, Section 7 presents oracle complexity results, and Section 8 provides the RA CLTs. Section 9 presents some numerical experiments along with insights and concluding remarks. The appendix provides the proof and an example related to an altenate complexity result and some statements from empirical process theory relevant to our RA setting.

2. MATHEMATICAL PRELIMINARIES   In what follows, we list notation and define key ideas that are repeatedly invoked in the paper.

2.1. Important Notation   (i) If $x \in \mathbb{R}^d$, then $x_i, i = 1, 2, \ldots, d$ refers to the i-th coordinate of x, implying that $x = (x_1, x_2, \ldots, x_d)$. (ii) For $x, y \in \mathbb{R}$, $a \vee b := \max(a, b)$ refers to the maximum of a and b, and $a \wedge b := \min(a, b)$ refers to the minimum of a and b. (iii) $\|x\|_p := \left(\sum_{i=1}^d |x_i|^p \right)^{\frac{1}{p}}$ refers to the L_p norm of $x \in \mathbb{R}^d$. As usual, we use the special notation $\|x\|$ for the $p = 2$ case to refer to the L_2 norm of $x \in \mathbb{R}^d$. (iv) For a random sequence $\{X_n, n \geq 1\}$, we write $X_n \overset{\text{wp1}}{\to} X$ to refer to almost sure convergence (also known as convergence with probability one), $X_n \overset{\text{p}}{\to} X$ to refer to convergence in probability, and $X_n \overset{\text{d}}{\to} X$ to refer to convergence in distribution (also known as weak convergence). (v) For positive valued sequences $\{a_n, n \geq 1\}$ and $\{b_n, n \geq 1\}$, we say $a_n \sim b_n$ to mean $a_n/b_n \to 1$ as $n \to \infty$. (vi) The function $\delta_{ij} = 1$ if $i = j$ and 0 otherwise refers to the Kronecker delta function.

2.2. Important Definitions

DEFINITION 1 (FIRST-ORDER CRITICAL POINTS).   The set $\mathscr{X}^*$ refers to the set of first-order critical points of the objective f, that is,

$$\mathscr{X}^* := \{x \in \mathbb{R}^d : \|\nabla f(x)\| = 0\}. \tag{1}$$

DEFINITION 2 (UNBIASED FIRST-ORDER STOCHASTIC ORACLE). We follow the oracle model detailed in [1, 49] within a sample-path context. So, by an unbiased stochastic first-order oracle associated with an objective function f, we are referring to a random map

$$(x, f, Y) \mapsto (F(x, Y), \nabla F(x, Y)),$$

where $\mathbb{E}[F(x,Y)] = f(x)$ and $\mathbb{E}[\nabla F(x,Y)] = \nabla f(x)$. An analogous definition holds when considering an unbiased first-order stochastic oracle within a filtered probability space.

DEFINITION 3 (GLOBALLY CONVERGENT). An iterative algorithm for identifying a first-order critical point of the function $f : \mathbb{R}^d \to \mathbb{R}$ is said to be *globally convergent* if, starting with any initial guess x_0, the sequence of iterates $\{x_n, n \geq 1\}$ generated by the algorithm has gradient norms converging to zero, that is,

$$\|\nabla f(x_n)\| \to 0 \text{ as } n \to \infty. \tag{2}$$

The iterative algorithm is *locally convergent* if (2) holds only when the initial guess x_0 is such that $\|\nabla f(x_0)\|$ is small enough.

DEFINITION 4 (LINEARLY AND SUB-LINEARLY CONVERGENT SOLVER). Suppose an iterative solver generates iterates $\{x_n, n \geq 1\}$ starting at the initial guess x_0 when identifying a first-order critical point of a function $f : \mathbb{R}^d \to \mathbb{R}$. Such a solver is said to exhibit (ρ, C)-*linear convergence* $(\rho < 1, C < \infty)$ if the gradient norm at the iterate x_n obtained after executing n steps of the algorithm satisfies

$$\|\nabla f(x_n)\| \leq C \|\nabla f(x_0)\| \rho^n. \tag{3}$$

A solver is said to exhibit $(\frac{1}{2}, C)$-*sublinear convergence* (with $C < \infty$) when executed on a function f bounded from below, that is, $f^* = \inf\{x \in \mathbb{R}^d : f(x)\} > -\infty$, if the gradient norm at the iterate x_n obtained after executing n steps of the algorithm satisfies

$$\|\nabla f(x_n)\| \leq C \left(f(x_0) - f(x_n)\right)^{1/2} n^{-1/2} \leq C \left(f(x_0) - f^*\right)^{1/2} n^{-1/2}. \tag{4}$$

REMARK 3. The definition in (4) expresses the progress made by a solver (in identifying a first-order critical point) in terms of the number of steps n and the initial optimality gap $f(x_0) - f^*$. To further understand the structure in (4), consider the gradient descent recursion

$$x_{k+1} = x_k - \eta \nabla f(x_k), k = 0, 1, \ldots; \quad x_n^* := \arg\min\{\|\nabla f(x_k)\| : 1 \leq k \leq n\}$$

when executed with step size $\eta \leq L^{-1}$ on an L-smooth, lower bounded function $f : \mathbb{R}^d \to \mathbb{R}$, $-\infty < f^* := \inf_{x \in \mathbb{R}^d} f(x)$. It is well-known [50, pp. 28] that the sequence $\{x_n^*, n \geq 1\}$ exhibits the global sub-linear convergence rate

$$\|\nabla f(x_n^*)\| \leq \left(\frac{2}{\eta}\right)^{1/2} (f(x_0) - f^*)^{1/2} n^{-1/2}, \tag{5}$$

thereby satisfying (4) with $C = \sqrt{\frac{2}{\eta}}$.

We present oracle complexity results based on (3) and (4).

DEFINITION 5 (EMPIRICAL MEASURE AND EMPIRICAL PROCESS). Let $\{Y_j, 1 \leq j \leq n\}$ be a sequence of independent and identically distributed (IID) copies of a random variable Y having distribution P on some appropriate measurable space $(\mathscr{Y}, \mathscr{A})$. The *empirical measure* P_n associated with $\{Y_j, j \geq 1\}$ is

$$P_n(A) := n^{-1} \sum_{j=1}^{n} \mathbb{I}_A(Y_j),$$

where $A \in \mathscr{A}$. Let $P(A) := P\{Y \in A\}$ which can also be written as $\mathbb{E}\mathbb{I}_A(Y)$. We define the *empirical process* as the following "scaled" version of the above:

$$G_n(A) := \sqrt{n}\left(P_n(A) - P(A)\right).$$

The above can also be defined for a general function F of Y:

$$P_n F := n^{-1} \sum_{j=1}^{n} F(Y_j), \quad G_n(F) := \sqrt{n}(P_n F - PF), \quad \text{where } PF = \mathbb{E}F(Y).$$

By the classical law of large numbers and central limit theorems, for any fixed $A \in \mathscr{A}$ and any fixed function F of Y, we have $P_n(A)$ and $P_n F$ converging almost surely to $P(A)$ and PF, and $G_n(A)$ and $G_n F$ converging in distribution to normal random variables with mean 0 and variances $P(A) - P(A)^2$ and $\mathbb{E}(F - PF)^2$. However, for many applications, we would like the above limit theorems to hold true *uniformly* over some class of subsets A and functions F.

The above concepts are very relevant to our current paper as our goal is exactly to minimize the empirical version of the random function F, with the ability to observe

$$\nabla f_n(x) := n^{-1} \sum_{j=1}^{n} \nabla F(x, Y_j) \; (= P_n \nabla F(x, \cdot)).$$

Note that now, instead of indexed by a set A, we are dealing with the class of functions indexed by $x \in \mathscr{R}^n$, $\{\nabla F(x, \cdot) : x \in \mathbb{R}^d\}$. Analogously, the "scaled" stochastic process

$$\left\{\sqrt{n}(\nabla f_n(x) - \nabla f(x)), x \in \mathbb{R}^d\right\} \quad \text{where } f(x) = \mathbb{E}F(x, Y)$$

is called the *empirical process*.

3. RA DESCRIPTION   In this section, we define RA more precisely and as it is applied to the type of stochastic optimization problems we consider in this paper.

Let $\{\tilde{Y}_k, k \geq 1\}$ be an IID ($\mathscr{Y}$-valued) random sequence on $(\Omega, \mathscr{F}, P)$. To precisely define a sequence of empirical mean function and gradient estimators, let's construct an increasing sequence of sigma algebras $\mathscr{F}_1 \subseteq \mathscr{F}_2 \subseteq \cdots$ as follows. Let M_1 be a constant, define

$$\begin{aligned} \mathscr{F}_1 &:= \sigma(\tilde{Y}_j | j \leq M_1); & M_2 &\in \mathscr{F}_1; \\ \mathscr{F}_2 &:= \sigma(\tilde{Y}_j | j \leq M_1 + M_2); & M_3 &\in \mathscr{F}_2; \\ &\;\;\vdots & & \\ \mathscr{F}_k &:= \sigma\left(\tilde{Y}_j \Big| j \leq \sum_{j=1}^{k} M_j\right); & M_{k+1} &\in \mathscr{F}_k. \end{aligned} \quad (6)$$

Let $\tilde{Y}$ be a copy identically distributed with $\{\tilde{Y}_k, k \geq 1\}$. Define the *sample-path function* f_m and the *sample-path gradient* having sample size m as follows:

$$f_m(x) := \frac{1}{m} \sum_{j=1}^{m} F(x, \tilde{Y}_j); \quad \nabla f_m(x) := \frac{1}{m} \sum_{j=1}^{m} \nabla F(x, \tilde{Y}_j) \quad x \in \mathbb{R}^d. \quad (7)$$

Thus we see that sample-path function $f_m(\cdot)$ and the sample-path gradient ∇f_m appearing in (7) are constructed as the sample mean of the random functions $F(\cdot, \tilde{Y}_j), j = 1, 2, \ldots, m$ and their gradients $\nabla F(\cdot, \tilde{Y}_j), j = 1, 2, \ldots, m$, respectively, where $\tilde{Y}_j, j = 1, 2, \ldots, m$ are identically distributed copies of

a random object $\tilde{Y}$. The random object $\tilde{Y}$ will never enter the stage of our presentation directly, but we assume that the gradients $\nabla F(\cdot, \tilde{Y})$ are unbiased with respect to $\nabla f(\cdot)$, that is,

$$\mathbb{E}\left[\nabla F(x, \tilde{Y})\right] = \nabla f(x), \quad x \in \mathbb{R}^d.$$

Due to the above construction of the filtration $(\Omega, \mathscr{F}, (\mathscr{F}_k)_{k \geq 1}, P)$, notions of a sample-path function and a sample-path gradient implicitly constructed with an adapted sample size $M_k \in \mathscr{F}_{k-1}$ are now well defined:

$$f_{M_k}(x) := \frac{1}{M_k} \sum_{j=1}^{M_k} F(x, Y_j); \quad \nabla f_{M_k}(x) := \frac{1}{M_k} \sum_{j=1}^{M_k} \nabla F(x, Y_j) \quad x \in \mathbb{R}^d, \tag{8}$$

where in (8), we have abused notation to obtain the convenient relabeling $Y_j := \tilde{Y}_{\left(\sum_{i=1}^{k-1} M_i\right)+j}$. While requiring elaborate filtration machinery for precise definition, (8) conveys the simple and intuitive idea that the function and gradient estimate at any point x during iteration k are each obtained as the sample mean of M_k unbiased observations. Furthermore, while we develop much of the theory that follows assuming that the sequence $\{\tilde{Y}_k, k \geq 1\}$ is IID, a relaxation of the the independence assumption is possible using the now standard machinery from empirical process theory [26, 75, 76]. Such relaxation will subsume important contexts where $\{\tilde{Y}_k, k \geq 1\}$ constitutes observations from a real system, e.g., a queuing network operating at equilibrium.

We assume throughout that we have at our disposal a *globally convergent* (see Definition 2) "iterative solver" that we generically call Method $\mathfrak{M}$, capable of solving the sample-path optimization problem:

$$\begin{aligned} &\text{minimize: } f_{M_k}(x) \\ &\text{subject to: } x \in \mathbb{R}^d. \end{aligned} \tag{Q_{M_k}}$$

More precisely, we assume having at our disposal Method $\mathfrak{M}$ which, given any $\epsilon > 0$ and starting with any initial guess, identifies a point $X_{M_k}(\epsilon) \in \mathbb{R}^d$ satisfying $\|\nabla f_{M_k}(X_{M_k}(\epsilon))\| \leq \epsilon$. Two examples of the Method $\mathfrak{M}$ that are of particular interest are line search and trust-region procedures [51] executed with L-BFGS [51] for Hessian estimation.

Next we give the listing of the RA Algorithm.

Algorithm 1 RA

1: **Input:** (i) initial guess X_0; (ii) procedure to update $\{M_k, k \geq 1\}$; (iii) procedure to update $\{\epsilon_k, k \geq 1\}$; (iv) Method $\mathfrak{M}$, e.g., line search or trust-region with L-BFGS.

2: **for** $k = 1, 2, \cdots$ **do**

3:     Set sample size and error tolerance: Choose $M_k \in \mathscr{F}_{k-1}$ and $\epsilon_k \in \mathscr{F}_{k-1}$.

4:     Execute Method $\mathfrak{M}$ for N_k steps: With Method $\mathfrak{M}$ and X_{k-1} as "warm-start," execute N_k steps to obtain 'inner iterates" $X_{k-1,t}, t = 0, 1, 2, \ldots, N_k$ satisfying

$$\|\nabla f_{M_k}(X_{k-1,N_k})\| \leq \epsilon_k, \quad \text{e.g., } \epsilon_k := \Gamma_k \|\nabla f_{M_k}(X_{k-1,0})\|, \text{ where } \Gamma_k > 0$$

5:     Set the k-th outer iterate: $X_k = X_{k-1,N_k}$.

6: **end for**

As seen in the algorithm listing, Algorithm RA is an iterative framework that is organized into outer and inner iterations. During the k-th (outer) iteration, Method $\mathfrak{M}$ uses the $(k-1)$-th

outer iterate as the initial guess to execute N_k steps and solve the sample-path problem (Q_{M_k}) to within adapted tolerance $\epsilon_k \in \mathscr{F}_{k-1}$. Let $\{X_k, k \geq 1\}$ denote the (outer) iterates, and $X_{k-1,t}, t = 0, 1, 2, \ldots, N_k$ the inner iterates generated by Method $\mathfrak{M}$ during iteration k. Since Method $\mathfrak{M}$ uses X_{k-1} as the initial guess during the k-th iteration and terminates after N_k steps when the norm of the sample-path gradient drops below ϵ_k, we see that the following hold:

$$X_{k-1,0} := X_{k-1}; \quad X_k := X_{k-1,N_k}; \quad \|\nabla f_{M_k}(X_k)\| \leq \epsilon_k.$$

We now briefly discuss four aspects of RA that are of vital theoretical and practical importance.

3.1. Choice of Method $\mathfrak{M}$. One of the key advantages of RA is that its design naturally allows "directly plugging-in" a highly successful deterministic optimization solver (generically called Method $\mathfrak{M}$ in Algorithm RA) in the service of solving a stochastic optimization problem. While Method $\mathfrak{M}$ can be any globally convergent procedure for identifying a first-order critical point in (Q_{M_k}), our particular interest is a quasi-Newton method such as line search or trust-region with L-BFGS [51] for Hessian estimation.

Step 4 in Algorithm RA assumes that the Method $\mathfrak{M}$ is able to solve (Q_{M_k}) to any prescribed level of tolerance ϵ_k, that is, Method $\mathfrak{M}$ is able to identify a point X_k such that $\|\nabla f_{M_k}(X_k)\| \leq \epsilon_k$. Such assumption is valid for standard quasi-Newton methods such as line search or trust-region [51] with L-BFGS [51] if the random functions $F(\cdot, Y)$ that comprise the sample-path function are $L(Y)$-smooth, that is,

$$\|\nabla F(x,Y) - \nabla F(y,Y)\| \leq L(Y)\|x - y\|, \quad \forall x, y \in \mathbb{R}^d, \quad \text{with } \mathbb{E}[L^2(Y)] < \infty.$$

It is well-known [35, 70] that when $F(\cdot, Y)$ is $L(Y)$-smooth and $\mathbb{E}[L^2(Y)] < \infty$, smoothness is transferred to the objective f, that is, the function $f(x) = \mathbb{E}[F(x,Y)], x \in \mathbb{R}^d$ is $\mathbb{E}[L]$-smooth.

3.2. Termination Criterion. Recall that RA executes Method $\mathfrak{M}$ on the sample-path problem (Q_{M_k}) during the k-th iteration. Since the sample-path problem (Q_{M_k}) is only a proxy for the true problem (Q), it stands to reason that Method $\mathfrak{M}$ not be executed ad infinitum but instead terminated if further solving is deemed to produce no statistical gains. RA formalizes this idea by introducing the notion of an error tolerance ϵ_k in Step 4 of RA: during the k-th iteration Method $\mathfrak{M}$ executes until it finds an iterate X_{k-1,N_k} satisfying $\|\nabla f_{M_k}(X_{k-1,N_k}\| \leq \epsilon_k$. While the choice of ϵ_k becomes a central question in our analysis, a common choice that is consistent with optimization practice is to choose $\epsilon_k := \Gamma_k \|\nabla f_{M_k}(X_{k-1,0})\| = \Gamma_k \|\nabla f_{M_k}(X_{k-1})\|$, implying that the Method $\mathfrak{M}$ executes as long as the norm of the sample gradient at the incumbent iterate remains above that at the initial guess by at least Γ_k. We calculate the oracle complexity resulting from this choice in Remark 6.

3.3. Warm Starts. RA employs "warm starts," that is, when executing Method $\mathfrak{M}$, the solution to the previous iteration becomes the initial guess for the subsequent iteration. The use of warm starts is more than just an implementation heuristic and features explicitly in the oracle complexity calculations of RA since they decide the number of inner iterations performed within RA. Specifically, warm starts reduce the number of inner iterations performed, and such reduction is featured prominently when the incumbent sample size is high.

3.4. Common Random Numbers The effectiveness of RA, particularly its use of a quasi-Newton solver to solve sample-path problems, is crucially dependent on the use of what is called *common random numbers* (CRN) [47]. Formally, notice that RA uses $Y_1, Y_2, \ldots, Y_{M_k}$ to generate the sample-path objective function f_{M_k}, that is, the "sample remains fixed" as the search procedure Method $\mathfrak{M}$ in Step 4 of RA is executed on $f_{M_k}(\cdot)$ for N_k steps towards solving the sample-path problem (Q_{M_k}). CRN should be seen as a mechanism that preserves any sample-path structure that may be present, e.g., strong convexity, in the process allowing faster convergence of the deterministic solver.

4. CONSISTENCY In this section, we present results that characterize the convergence behavior of the sequence $\{\|\nabla f(X_k)\|, k \geq 1\}$ of true gradient norms evaluated at the iterates generated by RA. To preserve generality and applicability, we will advocate that the relative error of the gradient estimator is a principal object dictating algorithmic behavior. More precisely, the *relative error class $\mathscr{R}$* is defined as:

$$\mathscr{R} := \{R(x, Y), x \in \mathbb{R}^d, Y \in \mathscr{Y}\}; \quad R(x, Y) := \frac{\nabla F(x, Y) - \nabla f(x)}{\|\nabla f(x)\| + c_0}; \quad c_0 > 0. \tag{9}$$

The constant c_0 is to avoid the nuisance associated with a vanishing gradient in the stationary set $\mathscr{X}^*$ (defined through (1)). The class $\mathscr{R}$ should be interpreted as the collection of the relative error functions $R(x, \cdot)$, labeled by $x \in \mathbb{R}^d$. The empirical version of the above is given by

$$R_m(x) := \frac{\nabla f_m(x) - \nabla f(x)}{\|\nabla f(x)\| + c_0}, \quad x \in \mathbb{R}^d, \quad \text{where} \quad \nabla f_m(x) := m^{-1} \sum_{j=1}^{m} \nabla F(x, Y_j). \tag{10}$$

In addition, we define the *envelope* and the *empirical envelope* as

$$G = G(Y) := \sup \left\{ \|R(x, Y)\| : x \in \mathbb{R}^d \right\}, \tag{11}$$

and

$$G_m(Y) := \sup\{\|R_m(x)\| : x \in \mathbb{R}^d\}. \tag{12}$$

Section 6 on example problem classes serves to support our claim that the relative error class affords great generality, and that *relative error* (and not absolute error) of the gradient estimator is a fundamental object that decides consistency and convergence rates in stochastic optimization. The simple but widely used multiplicative error class, appearing as Example II in Section 6, illustrates this point well — it satisfies stipulations on the uniform law of the relative error for strong consistency but violates the corresponding routinely assumed requirements for consistency on the absolute error.

4.1. Main Results on Consistency In what follows we demonstrate RA's strong consistency property by combining conditions on the behavior of the termination criterion sequence $\{\epsilon_k, k \geq 1\}$ and the sample size sequence $\{M_k, k \geq 1\}$ with a uniform law on the relative error class.

THEOREM 1 **(Strong Consistency of RA).** *Suppose that the relative error class $\mathscr{R}$ defined in (9) satisfies the uniform law*

$$\sup_{x \in \mathbb{R}^d} \|R_m(x)\| \stackrel{wp1}{\to} 0 \ as \ m \to \infty, \tag{R-ULLN}$$

and that the RA's termination criterion sequence $\{\epsilon_k, k \geq 1\}, \epsilon_k \in \mathscr{F}_{k-1}$ satisfies

$$\epsilon_k \stackrel{wp1}{\to} 0 \ as \ k \to \infty. \tag{13}$$

(a) *If the envelope G (defined in (11)) of the relative error class $\mathscr{R}$ is such that $E[G] < \infty$ and the sample size $M_k \stackrel{wp1}{\to} \infty$, then*

$$\|\nabla f(X_k)\| \stackrel{p}{\to} 0 \ as \ k \to \infty. \tag{14}$$

(b) *If the envelope G is such that $E[G^2] < \infty$ and the sample sizes satisfy $\sum_{k=1}^{\infty} M_k^{-1} < \infty \ wp1$, then*

$$\|\nabla f(X_k)\| \stackrel{wp1}{\to} 0 \ as \ k \to \infty. \tag{15}$$

Theorem 1 gives sufficient conditions for consistency in probability and with probability one. A third type of consistency that is more routinely employed in the optimization literature is what we call L_1 consistency, that is,

$$\mathbb{E}\left[\|\nabla f(X_k)\| \,|\, \mathscr{F}_{k-1}\right] \to 0.$$

As will become evident shortly through the proof of Theorem 2, in addition to the uniform convergence (R-ULLN) of the relative error class $\mathscr{R}$, L_1 consistency follows if the iterates are such that $\mathbb{E}\left[\|\nabla f(X_k)\|^2 \,|\, \mathscr{F}_{k-1}\right]$ remains bounded, implying in effect that the variance in the iterates is not so high that the norm of the true gradient evaluated at the iterates doesn't diverge in expectation. As the examples in Section 6 will illustrate, our use of relative error demonstrates that this condition is widely satisfied. Similar to Theorem 1(a) for convergence in probability, Theorem 2 also imposes the condition $\epsilon_k \to 0$ on the error tolerance sequence $\{\epsilon_k, k \geq 1\}$; this is in contrast to Theorem 1(b) for almost sure convergence, which imposes the slightly more stringent condition $\sum_k M_k^{-1} < \infty$ on the growth of the sample size sequence $\{M_k, k \geq 1\}$.

THEOREM 2 (L_1 **Consistency of RA**). *Suppose the sample size sequence $\{M_k, k \geq 1\}$ and the error-tolerance sequence $\{\epsilon_k, k \geq 1\}$ are chosen so that*

$$M_k \overset{wp1}{\to} \infty; \quad \mathbb{E}\left[\epsilon_k \,|\, \mathscr{F}_{k-1}\right] \to 0,$$

and the relative error class $\mathscr{R}$ defined in (9) *satisfies the uniform law in* (R-ULLN) *along with* $\mathbb{E}[G^2(Y)] < \infty$. *Suppose also that the iterates $\{X_k, k \geq 1\}$ are such that*

$$\sup_k \mathbb{E}[\|\nabla f(X_k)\|^2 \,|\, \mathscr{F}_{k-1}] =: \ell_{2,\infty}^2 < \infty. \tag{Bd-Grad}$$

Then as $k \to \infty$,

$$\mathbb{E}\left[\|\nabla f(X_k)\| \,|\, \mathscr{F}_{k-1}\right] \to 0. \tag{16}$$

REMARK 4. Theorem 1 prominently assumes a uniform law on the relative error class. As we shall see later, verifying this condition is fairly direct in the context of optimization, in particular for explicit examples. On the other hand, it is instructive to relate this condition to a more fundamental property about the underlying stochasticity of the oracle. Qualitatively, this relation touches upon the "regularity" or "irregularity" of the stochastic functions $F(x, Y)$ used by the oracle to approximate the function f to be optimized. This has been studied extensively in the field of empirical processes. The quantitative measure is given by the entropy of the class of functions $\{F(x, Y), x \in \mathscr{R}^d\}$. Intuitively, if the entropy does not grow too fast as the fineness of approximation increases, the approximating functions f_n can represent the true function f more faithfully so that minimizing f_n can give a more trustworthy answer. In fact, there is a more or less equivalence between the uniform law and the sub-linear growth of the entropy function. We refer to [26, 75, 76] for comprehensive theory. Some of the key results will be illustrated in the appendix.

4.2. Proofs of Theorem 1 and Theorem 2.
Proof of Theorem 1. Since (R-ULLN) holds along with $M_k \overset{\mathrm{wp1}}{\to} \infty$, we have

$$\lim_{k\to\infty} P\left(\sup_{x\in\mathbb{R}^d}\|R_{M_k}(x)\| > \epsilon \,|\, \mathscr{F}_{k-1}\right) = 0 \quad \forall \epsilon > 0, \tag{17}$$

and hence (using bounded convergence theorem [7]), for all $\epsilon > 0$,

$$\lim_{k\to\infty} P\left(\sup_{x\in\mathbb{R}^d}\|R_{M_k}(x)\| > \epsilon\right) = \lim_{k\to\infty} \mathbb{E}\left[P\left(\sup_{x\in\mathbb{R}^d}\|R_{M_k}(x)\| > \epsilon \,|\, \mathscr{F}_{k-1}\right)\right]$$
$$= \mathbb{E}\left[\lim_{k\to\infty} P\left(\sup_{x\in\mathbb{R}^d}\|R_{M_k}(x)\| > \epsilon \,|\, \mathscr{F}_{k-1}\right)\right] = 0. \tag{18}$$

Define the following events for convenience.

$$\mathcal{B}_{0,k} := \{\|\nabla f(X_k)\| > t\}, \quad t \in (0,1);$$
$$\mathcal{B}_{1,k} := \left\{\|\nabla f_{M_k}(X_k) - \nabla f(X_k)\| \geq \tilde{t}(\|\nabla f(X_k)\| + c_0)\right\},$$

where

$$\tilde{t} := \frac{t}{2c_0 + t} = 1 - \frac{2c_0}{2c_0 + t}.$$

We have

$$
\begin{aligned}
P\left(\mathcal{B}_{0,k}\right) &= P\left(\mathcal{B}_{0,k} \bigcap \mathcal{B}_{1,k}\right) + P\left(\mathcal{B}_{0,k} \bigcap \mathcal{B}_{1,k}^c\right) \\
&\leq P\left(\mathcal{B}_{1,k}\right) + P\left(\mathcal{B}_{0,k} \bigcap (1-\tilde{t})\|\nabla f(X_k)\| \leq \|\nabla f_{M_k}(X_k)\| + \tilde{t}c_0\right) \\
&\leq P\left(\mathcal{B}_{1,k}\right) + P\left(\mathcal{B}_{0,k} \bigcap \|\nabla f(X_k)\| \leq \frac{\epsilon_k + \tilde{t}c_0}{1-\tilde{t}}\right) \\
&= \underbrace{P\left(\mathcal{B}_{1,k}\right)}_{I_1} + \underbrace{P\left(\mathcal{B}_{0,k} \bigcap \|\nabla f(X_k)\| \leq \epsilon_k\left(1 + \frac{t}{2c_0}\right) + \frac{t}{2}\right)}_{I_2}.
\end{aligned}
\tag{19}
$$

Now notice that

$$
\begin{aligned}
\mathcal{B}_{0,k} := \{\|\nabla f(X_k)\| > t\} &\subseteq \left\{(1-\tilde{t})\|\nabla f(X_k)\| > \tilde{t}c_0 + \epsilon_k\right\} \\
&\subseteq \left\{(1-\tilde{t})\|\nabla f(X_k)\| > \tilde{t}c_0 + \|\nabla f_{M_k}(X_k)\|\right\} \\
&\subseteq \left\{\|\nabla f_{M_k}(X_k) - \nabla f(X_k)\| > \tilde{t}(\|\nabla f(X_k)\| + c_0)\right\} \\
&\equiv \left\{\|R_{M_k}(X_k)\| > \tilde{t}\right\} =: \mathcal{B}_{1,k}.
\end{aligned}
\tag{20}
$$

From (20) and (18), we see that $I_1 \to 0$ as $k \to \infty$; and $I_2 = 0$ for large enough k by observing the definition of $\mathcal{B}_{0,k}$ and since $\epsilon_k \overset{\mathrm{wp1}}{\to} 0$ as $k \to \infty$. Conclude that (14) holds.

To prove the assertion in (15), we will first show that

$$\sum_{k=1}^{\infty} P\left(\sup_{x \in \mathbb{R}^d} \|R_{M_k}(x)\| > \epsilon \,\big|\, \mathscr{F}_{k-1}\right) < \infty. \tag{21}$$

Define the restricted class

$$\mathscr{R}_u := \{R(x, Y)\,\mathbb{I}(G(Y) \leq u) : x \in \mathbb{R}^d\}, \quad u > 0 \tag{22}$$

and, for ease of exposition, the following two norms associated with the empirical average of the relative errors on the classes $\mathscr{R}$ and $\mathscr{R}_u$.

$$
\begin{aligned}
\|R_{M_k}\|_{\mathscr{R}} &:= \sup_{x \in \mathbb{R}^d} \left\{\left\|\frac{1}{M_k}\sum_{j=1}^{M_k} R(x, Y_j)\right\|\right\}; \\
\|R_{M_k}\|_{\mathscr{R}_u} &:= \sup_{x \in \mathbb{R}^d} \left\{\left\|\frac{1}{M_k}\sum_{j=1}^{M_k} R(x, Y_j)\,I(G(Y_j) \leq u)\right\|\right\}
\end{aligned}
\tag{23}
$$

Due to (R-ULLN) and since $\mathscr{R}_u$ is a bounded class of random variables $\mathbb{E}\left[\|R_{M_k}\|_{\mathscr{R}_u}\,|\,\mathscr{F}_{k-1}\right] \leq \epsilon$ for k large enough, and arbitrary $\epsilon > 0$.

Define $A_{k,u} := \mathbb{E}\left[\|R_{M_k}\|_{\mathscr{R}_u}\,|\,\mathscr{F}_{k-1}\right]$ and observe that for large enough k,

$$P\left(\|R_{M_k}\|_{\mathscr{R}} > t \,|\, \mathscr{F}_{k-1}\right) \leq P\left(\left|\|R_{M_k}\|_{\mathscr{R}} - A_{k,u}\right| > (t - \epsilon) \,|\, \mathscr{F}_{k-1}\right). \tag{24}$$

To bound the right-hand side of (24), write

$$
\begin{aligned}
&P\left(\left|\|R_{M_k}\|_{\mathscr{R}}-A_{k,u}\right|>t-\epsilon\,|\,\mathscr{F}_{k-1}\right)\\
&\quad=P\left(\left|\sup_{x\in\mathbb{R}^d}\left\|\frac{1}{M_k}\sum_{j=1}^{M_k}R(x,Y_j)\left(\mathbb{I}(G(Y_j)\le u)+\mathbb{I}(G(Y_j)>u)\right)\right\|-\mathbb{E}\left[\|R_{M_k}\|_{\mathscr{R}_u}\,|\,\mathscr{F}_{k-1}\right]\right|>t-\epsilon\,|\,\mathscr{F}_{k-1}\right)\\
&\quad\le P\left(\left|\|R_{M_k}\|_{\mathscr{R}_u}-\mathbb{E}\left[\|R_{M_k}\|_{\mathscr{R}_u}\,|\,\mathscr{F}_{k-1}\right]\right|>\frac{t-\epsilon}{2}\,|\,\mathscr{F}_{k-1}\right)\\
&\qquad\qquad\qquad\qquad\qquad\qquad+P\left(\sup_{x\in\mathbb{R}^d}\left\|\frac{1}{M_k}\sum_{j=1}^{M_k}R(x,Y_j)\mathbb{I}(G(Y_j)>u)\right\|>\frac{t-\epsilon}{2}\,|\,\mathscr{F}_{k-1}\right)\\
&\quad\le\underbrace{P\left(\left|\|R_{M_k}\|_{\mathscr{R}_u}-\mathbb{E}\left[\|R_{M_k}\|_{\mathscr{R}_u}\right]\right|>\frac{t-\epsilon}{2}\,|\,\mathscr{F}_{k-1}\right)}_{T_{1,k}}+\underbrace{P\left(\frac{1}{M_k}\sum_{j=1}^{M_k}G(Y_j)>\frac{t-\epsilon}{2}\,|\,\mathscr{F}_{k-1}\right)}_{T_{2,k}}. \quad (25)
\end{aligned}
$$

Let's now analyze each of the terms $T_{1,k},T_{2,k}$ appearing in (25). From Talagrand's inequality [26, pp. 158],

$$
T_{1,k}\le\kappa\exp\left\{-\frac{1}{\kappa}\frac{M_k(t-\epsilon)}{u}\log(1+\frac{(t-\epsilon)u}{v})\right\}, \quad (26)
$$

where κ is a universal constant, and $v<\infty$ is such that

$$
\mathbb{E}\left[\sup_{x\in\mathbb{R}^d}\left\{\left\|\frac{1}{M_k}\sum_{j=1}^{M_k}R^2(x,Y_j)\right\|\right\}\right]\le v.
$$

(Such a v is guaranteed to exist since we have assumed that the envelope G has finite second moment.) Also, we know from Markov's inequality [7, pp. 276] that

$$
T_{2,k}=P\left(\frac{1}{M_k}\sum_{j=1}^{M_k}G(Y_j)>\frac{t-\epsilon}{2}\,|\,\mathscr{F}_{k-1}\right)\le\frac{4}{(t-\epsilon)^2}\frac{\mathrm{Var}\left[G\,\mathbb{I}(G>u)\,|\,\mathscr{F}_{k-1}\right]}{M_k}, \quad (27)
$$

where the numerator of the right-hand side of (27) is finite because part (b) assumes that $\mathbb{E}[G^2]<\infty$. From (26), (27) and the assumed lower bound $\sum_k M_k^{-1}<\infty$ on sample sizes, we see that

$$
\sum_{k=1}^{\infty}P\left(\left|\|R_{M_k}\|_{\mathscr{R}}-\mathbb{E}\left[\|R_{M_k}\|_{\mathscr{R}_u}\,|\,\mathscr{F}_{k-1}\right]\right|>t-\epsilon\,|\,\mathscr{F}_{k-1}\right)\le\sum_{k=1}^{\infty}(T_{1,k}+T_{2,k})<\infty. \quad (28)
$$

From (24) and (28) we see that

$$
\sum_{k=1}^{\infty}P\left(\|R_{M_k}\|_{\mathscr{R}}>t\,|\,\mathscr{F}_{k-1}\right)<\infty, \quad (29)
$$

and (21) holds. To complete the proof of the assertion in (15), we again use (19); since there exists $k_0<\infty$ such that $I_2=0$ for $k\ge k_0$, we can write

$$
\sum_{k=1}^{\infty}P\left(\mathscr{B}_{0,k}\,|\,\mathscr{F}_{k-1}\right)=\sum_{k=1}^{\infty}P\left(\mathscr{B}_{1,k}\,|\,\mathscr{F}_{k-1}\right)+\sum_{k=1}^{k_0}P\left(\mathscr{B}_{0,k}\bigcap\|\nabla f(X_k)\|\le\epsilon_k+\frac{t}{2}\,|\,\mathscr{F}_{k-1}\right). \quad (30)
$$

Now use (21) and (30) to see that

$$
\sum_{k=1}^{\infty}P\left(\mathscr{B}_{1,k}\,|\,\mathscr{F}_{k-1}\right)<\infty, \quad (31)
$$

implying that (15) holds from the conditional version of Borel-Cantelli [80]. ∎

Proof of Theorem 2. Recall the empirical-mean relative error $R_m(x)$ and the empirical-mean envelope as defined in (10) and (12). We see after some algebra that

$$
\begin{aligned}
\|\nabla f(X_k)\| &\leq \|\nabla f_{M_k}(X_k)\| + \|\nabla f_{M_k}(X_k) - \nabla f(X_k)\| \\
&\leq \|\nabla f_{M_k}(X_k)\| + R_{M_k}(X_k)(c_0 + \|\nabla f(X_k)\|) \\
&\leq \|\nabla f_{M_k}(X_k)\| + c_0 G_{M_k} + G_{M_k}\|\nabla f(X_k)\|
\end{aligned}
\tag{32}
$$

Recalling that $\|\nabla f_{M_k}(X_k)\| \leq \epsilon_k$, taking conditional expectations on (32), and then applying Cauchy-Schwarz inequality, we get

$$
\begin{aligned}
\mathbb{E}\left[\|\nabla f(X_k)\| \mid \mathscr{F}_{k-1}\right] &\leq \mathbb{E}[\epsilon_k \mid \mathscr{F}_{k-1}] + c_0 \mathbb{E}\left[G_{M_k} \mid \mathscr{F}_{k-1}\right] + \mathbb{E}\left[G_{M_k}\|\nabla f(X_k)\| \mid \mathscr{F}_{k-1}\right] \\
&\leq \mathbb{E}[\epsilon_k \mid \mathscr{F}_{k-1}] + c_0 \mathbb{E}\left[G_{M_k} \mid \mathscr{F}_{k-1}\right] + \left(\mathbb{E}\left[G_{M_k}^2 \mid \mathscr{F}_{k-1}\right]\right)^{1/2} \left(\mathbb{E}\left[\|\nabla f(X_k)\|^2 \mid \mathscr{F}_{k-1}\right]\right)^{1/2} \\
&\leq \mathbb{E}[\epsilon_k \mid \mathscr{F}_{k-1}] + c_0 \mathbb{E}\left[G_{M_k} \mid \mathscr{F}_{k-1}\right] + \left(\mathbb{E}\left[G_{M_k}^2 \mid \mathscr{F}_{k-1}\right]\right)^{1/2} \ell_{2,\infty},
\end{aligned}
\tag{33}
$$

where the last inequality in (33) follows from (Bd-Grad). However, since we have assumed that $\mathbb{E}[G^2(Y)] < \infty$ and that (R-ULLN) hold, it follows that

$$
\mathbb{E}\left[G_{M_k}^2 \mid \mathscr{F}_{k-1}\right] \to 0 \text{ and } \mathbb{E}\left[G_{M_k} \mid \mathscr{F}_{k-1}\right] \to 0.
\tag{34}
$$

Using (34) and the assumption $\mathbb{E}\left[\epsilon_k \mid \mathscr{F}_{k-1}\right] \to 0$ in (33), we see that the assertion in (16) holds. ∎

5. ITERATION COMPLEXITY   In this section, under a further strengthening of the uniform law assumption in (R-ULLN), we derive guarantees on RA's iteration complexity along with some useful corollaries on ϵ-complexity.

It should be clear that both the error tolerance sequence $\{\epsilon_k, k \geq 1\}$ and the sample size sequence $\{M_k, k \geq 1\}$ interact to decide the iteration complexity. To understand these effects better, for all results that follow, we assume a power relationship between the error tolerance and the sample size. Precisely, we assume that for each $k \geq 1$,

$$
\epsilon_k := \frac{C_{2,k}}{M_k^p}, p > 0 \quad \text{and} \quad c_2 \geq C_{2,k} \in \mathscr{F}_{k-1},
\tag{C.2(a)}
$$

where p is a constant that we will argue later should be set to $1/2$, and $C_{2,k}$ is an $\mathscr{F}_{k-1}$-measurable random variable with the upper bound c_2. We now state the main iteration complexity result.

THEOREM 3 **(Iteration Complexity)**.   *Let the postulates of Theorem 2 hold. Furthermore, let the empirical-mean envelope*

$$
G_m := \sup_{x \in \mathscr{X}} \|R_m(x)\|
$$

exhibit "CLT-scaling", that is, there exists $\sigma < \infty$ such that,

$$
\mathbb{E}\left[\left(\sqrt{M_k}\, G_{M_k}\right)^2 \mid \mathscr{F}_{k-1}\right] \leq \sigma^2; \quad M_k \in \mathscr{F}_{k-1}.
\tag{CLT-sc}
$$

Then, for $k \geq 1$ we have that

$$
\mathbb{E}\left[\|\nabla f(X_k)\| \mid \mathscr{F}_{k-1}\right] \leq \frac{c_2}{M_k^p} + \frac{(c_0\sigma + \ell_{2,\infty}\sigma)}{\sqrt{M_k}}.
\tag{35}
$$

Choose the sample-size sequence $\{M_k, k \geq 1\}$ so that $M_1 = m_1$ and

$$
M_k := C_{1,k}M_{k-1}; \quad c_1 \leq C_{1,k} \in \mathscr{F}_{k-1},
\tag{C.1(a)}
$$

and the error tolerance as in (C.2(a)) with $p = 1/2$. Then for $k \geq 1$,

$$
\mathbb{E}\left[\|\nabla f(X_k)\|\right] \leq \left(\frac{1}{\sqrt{c_1}}\right)^k \left(\frac{(c_2 + c_0\sigma + \ell_{2,\infty}\sigma)}{\sqrt{m_1}}\right).
\tag{36}
$$

We see that in addition to a uniform law on the relative error class that is needed for consistency of RA, iteration complexity entails a certain regularity of the envelope, expressed through the CLT-scaling assumption (CLT-sc). This condition, loosely speaking, stipulates that the supremum norm of the relative error scales as $O(1/\sqrt{M_k})$ in expectation. Similar to the description in Remark 4, such a CLT-scaling is related to some integrability condition of the entropy function. This condition can be expressed concisely in terms of the Koltchinskii-Pollard entropy function. See [26, Theorems 3.5.1 and 3.5.4] and also the appendix for a brief exposition. We expect the assumption in (CLT-sc) to hold widely, in particular for the examples in Section 6 as it is made on the relative error envelope rather than the absolute error envelope.

Recall that Theorem 3 assumes the relationship $\epsilon_k = C_{2,k}/M_k^p$ between the error tolerance and the sample size during each iteration, where the exponent p is to be chosen. By observing the derived complexity in (35), we deduce that the fastest achievable iteration complexity by RA is $O(1/M_k^{p\wedge 1/2})$ suggesting the choice $p \geq 1/2$. We can further reason that one should choose $p = 1/2$ since choosing $p > 1/2$ amounts to "oversolving," that is, leads to increased computational effort during each iteration k without yielding a better complexity rate. We will make this idea of not "oversolving" more precise in Section 7, where we present the oracle complexity of RA.

Theorem 3 also leads to a simple but useful corollary that characterizes an upper bound on the number of iterations needed for ϵ-complexity.

COROLLARY 1 (ϵ-**complexity of RA**). *Let the postulates Theorem 3 along with the sample size choice in* (C.1(a)) *and the error tolerance choice in* (C.2(a)) *with $p = 1/2$. Then for all $k \geq K(\epsilon)$ where*

$$K(\epsilon) := 1 + \frac{1}{\log\sqrt{c_1}} \left(\log\frac{1}{\epsilon} + \log\frac{c_2 + c_0\sigma + \ell_{2,\infty}\sigma}{\sqrt{m_1}} \right),$$

we have that

$$\mathbb{E}\left[\|\nabla f(X_k)\| \mid \mathscr{F}_{k-1}\right] \leq \epsilon, \quad \epsilon > 0.$$

Proof of Theorem 3. From (33), we have

$$\mathbb{E}\left[\|\nabla f(X_k)\| \mid \mathscr{F}_{k-1}\right] \leq \mathbb{E}[\epsilon_k \mid \mathscr{F}_{k-1}] + c_0\mathbb{E}\left[G_{M_k} \mid \mathscr{F}_{k-1}\right] + \mathbb{E}\left[G_{M_k}\|\nabla f(X_k)\| \mid \mathscr{F}_{k-1}\right]$$
$$\leq \mathbb{E}[\epsilon_k \mid \mathscr{F}_{k-1}] + c_0\mathbb{E}\left[G_{M_k} \mid \mathscr{F}_{k-1}\right] + \left(\mathbb{E}\left[G_{M_k}^2 \mid \mathscr{F}_{k-1}\right]\right)^{1/2}\left(\mathbb{E}\left[\|\nabla f(X_k)\|^2 \mid \mathscr{F}_{k-1}\right]\right)^{1/2}.$$
$$(37)$$

From (Bd-Grad), we know that there exists $0 < \ell_{2,\infty} < \infty$ such that

$$\mathbb{E}\left[\|\nabla f(X_k)\|^2 \mid \mathscr{F}_{k-1}\right] \leq \ell_{2,\infty}^2 < \infty. \tag{38}$$

Use (CLT-sc), (C.2(a)), and (38) to see that the assertion in (35) holds. To obtain (36), recurse the upperbound in (35) backward with $p = 1/2$. ∎

In the following we state an alternate iteration complexity result that dispenses with the bounded iterates assumption in (Bd-Grad) needed for L_1 consistency, while strengthening the assumption in (R-ULLN) using the notion of a *weighted relative error class*.

THEOREM 4 (**Alternate Iteration Complexity**). *Define, for $\alpha \in (0,1)$ and $\delta_0 > 0$, the (weighted) relative error class*

$$w\mathscr{R} := \left\{ R(x,Y)\left(\|\nabla f(x)\|^\alpha + \delta_0\right), x \in \mathbb{R}^d \right\}; \quad R(x,Y) := \frac{\nabla F(x,Y) - \nabla f(x)}{\|\nabla f(x)\| + c_0} \tag{39}$$

and its empirical-mean envelope

$$H_m := \sup_{x \in \mathscr{X}} \|R_m(x)\|\left(\|\nabla f(x)\|^\alpha + \delta_0\right).$$

Suppose, for some $\alpha > 0$ and $\delta_0 > 0$, the empirical-mean envelope of the weighted relative error class exhibits "CLT-scaling," that is there exists $\sigma_0 < \infty$ such that,

$$\mathbb{E}\left[\left(\sqrt{M_k}\, H_{M_k}\right)^{1/\alpha} \mid \mathscr{F}_{k-1}\right] \leq \sigma_0^{1/\alpha}; \quad M_k \in \mathscr{F}_{k-1}. \tag{40}$$

Then, for $k \geq 2$ and $\kappa_2 := 2(1-\alpha)\left(\frac{2\alpha}{1-\alpha}\right)^{\alpha}\sigma_0$, we have that

$$\mathbb{E}\left[\|\nabla f(X_k)\| \mid \mathscr{F}_{k-1}\right] \leq \frac{2c_2}{M_k^p} + \frac{2c_0\sigma_0/\delta_0}{\sqrt{M_k}} + \left(\frac{\kappa_2}{\sqrt{M_k}}\right)^{1/\alpha}. \tag{41}$$

Choose the sample-size sequence $\{M_k, k \geq 1\}$ according to (C.1(a)), *and error tolerance according to* (C.2(a)) *with $p = 1/2$. Then for $k \geq 2$,*

$$\mathbb{E}\left[\|\nabla f(X_k)\|\right] \leq \left(\frac{1}{\sqrt{c_1}}\right)^{k-1}\left(\frac{2(c_2 + c_0\sigma_0/\delta_0)}{\sqrt{m_1}} + \left(\frac{\kappa_2}{\sqrt{m_1}}\right)^{1/\alpha}\right). \tag{42}$$

Analogous to Corollary 1, Theorem 4 leads to the following simple corollary that provides an ϵ-complexity result.

COROLLARY 2 (ϵ-**complexity of Alternate RA**). *Let the postulates Theorem 4 along with the sample size choice in* (C.1(a)) *and the error tolerance choice in* (C.2(a)) *with $p = 1/2$. Then for all $k \geq K(\epsilon)$ where*

$$K(\epsilon) := 1 + \frac{1}{\log\sqrt{c_1}}\left(\log\frac{1}{\epsilon} + \log\left(\frac{2(c_2 + c_0\sigma_0/\delta_0)}{\sqrt{m_1}} + \left(\frac{\kappa_2}{\sqrt{m_1}}\right)^{1/\alpha}\right)\right),$$

we have that

$$\mathbb{E}\left[\|\nabla f(X_k)\| \mid \mathscr{F}_{k-1}\right] \leq \epsilon, \quad \epsilon > 0.$$

The proof of Theorem 4 will be given in the Appendix.

6. EXAMPLE PROBLEM CLASSES It should be clear that the uniform law (R-ULLN) on the relative error class (or the sub-linear entropy growth condition in (Sub-Ent), as detailed in the Appendix) and the bounded gradients assumption appearing in (Bd-Grad) essentially ensure in-probability, almost-sure, and L_1 convergence of RA's iterates. How stringent are (R-ULLN) and (Bd-Grad)? To further the reader's intuition in support of our claim that (R-ULLN) and (Bd-Grad) are widely satisfied, we now describe a number of example problem classes, in each case discussing whether (R-ULLN) and (Bd-Grad) are satisfied.

6.1. Example (Uniform Error Class). Suppose that the error in the gradient norm is uniformly convergent, that is,

$$\sup_{x\in\mathscr{X}}\|\nabla f_m(x) - \nabla f(x)\| \overset{\text{wp1}}{\to} 0 \text{ as } m \to \infty, \tag{ULLN}$$

and that the envelope

$$H = H(Y) := \sup\{\|\nabla F(x,Y) - \nabla f(x)\| : x \in \mathscr{X}\}$$

has finite second moment, that is,

$$\mathbb{E}[H^2] < \infty. \tag{43}$$

Variations of such a uniform error assumption are standard in SAA and empirical risk minimization literature, e.g., see [70, pp. 157,174] and [4, 12, 44, 45].) Under (ULLN) and (43), notice that

$$\sup_{x \in \mathscr{X}} \|R_m(x)\| = \sup_{x \in \mathscr{X}} \frac{\|\nabla f_m(x) - \nabla f(x)\|}{\|\nabla f(x)\| + c_0} \leq \sup_{x \in \mathscr{X}} \frac{\|\nabla f_m(x) - \nabla f(x)\|}{c_0} \overset{\text{wp1}}{\to} 0, \tag{44}$$

implying that (R-ULLN) is satisfied, and that strong consistency is guaranteed if RA is implemented on this class with $M_k \to \infty$ and $\epsilon_k \to 0$.

Let's next demonstrate that the condition in (Bd-Grad) needed for L_1 consistency of RA's iterates is satisfied as well. First, note that due to the assumption in (ULLN), the δ-covering number of the absolute error class

$$\mathscr{A} := \{\nabla F(x, Y) - \nabla f(x), x \in \mathscr{X}\}$$

exhibits sub-linear empirical entropy growth, that is,

$$\frac{1}{M_k} H_1(\delta, \mathscr{A}, P_{M_k}) \overset{\text{p}}{\to} 0 \text{ as } M_k \to \infty. \tag{45}$$

Due to (45), we see that the postulates of Theorem 2.4.3 [76, pp. 123] hold, and we have

$$\mathbb{E}\left[\sup_{x \in \mathscr{X}} \|\nabla f_{M_k}(x) - \nabla f(x)\| \,\big|\, \mathscr{F}_{k-1}\right] \to 0 \text{ as } k \to \infty. \tag{46}$$

Recall that the iterates $\{X_k, k \geq 1\}$ generated by RA are such that

$$\|\nabla f_{M_k}(X_k)\| \leq \epsilon_k. \tag{47}$$

Using (46) and (47), and since $\epsilon_k \to 0$, we conclude that the key condition in (Bd-Grad) holds. ∎

Indeed, while (ULLN) is a common assumption, it is too stringent and is frequently violated by even standard stochastic optimization settings. Let's illustrate this with a simple example. Suppose $\mathscr{X} = \mathbb{R}^d$, $b \in \mathbb{R}$, $a = \mathbb{E}[Z(Y)]$, and $Z(Y) > 0$ is some positive-valued random variable with finite expectation. Set

$$F(x, Y) = \frac{1}{2} Z(Y) x^T x + bx; \quad f(x) = \frac{1}{2} a x^T x + bx. \tag{48}$$

We then see that

$$\nabla F(x, Y) = Z(Y)x + b; \quad \nabla f(x) = ax + b$$

and hence, if $\text{Var}(Z(Y)) \neq 0$, then

$$\sup_{x \in \mathscr{R}^d} \|\nabla f_m(x) - \nabla f(x)\| = \left|\frac{1}{n} \sum_{j=1}^{n} Z(Y_j) - a\right| \sup_{x \in \mathscr{R}^d} \|x\| \not\to 0,$$

thus violating (ULLN).

The main feature of the counter-example in (48) responsible for (ULLN) violation is that the error in the sampled gradient is proportional to the true gradient, that is,

$$\|\nabla F(x, Y) - \nabla f(x)\| = \|(Z(Y) - a)x\| \geq \left|\frac{Z(Y) - a}{a}\right| (\|\nabla f(x)\| - b). \tag{49}$$

The inequality in (49) implies that the absolute error in the gradient estimate for the counter-example in (48) will increase proportionally with the norm of the true gradient, thereby violating (ULLN) if the true gradient is unbounded over $\mathscr{X}$.

This issue of absolute error in the estimated gradient norm being proportional to the true gradient seems to be common in practice [14]. Through the next example class, we demonstrate that in many simple settings, while a uniform law (ULLN) on the absolute error is violated, a uniform law (R-ULLN) on the relative error is satisfied.

6.2. Example (Multiplicative Error Class). Suppose the random function $\nabla F(\cdot, Y)$ is such that

$$\nabla F(x, Y) = A(Y)\nabla f(x) + B(Y), \quad x \in \mathbb{R}^d \tag{50}$$

where $A(Y)$ is a $d \times d$ matrix of real-valued random variables and $B(Y)$ is an $\mathbb{R}^d$ valued random variable satisfying

$$\mathbb{E}[A(Y)] = J_d; \quad \mathbb{E}[B(Y)] = 0, \tag{51}$$

and J_d is the $d \times d$ matrix of ones. Then, we see that

$$R_m(x) = \frac{(\bar{A}_m - J_d)\nabla f(x) + \bar{B}_m}{\|\nabla f(x)\| + c_0}; \quad \bar{A}_m := \frac{1}{m}\sum_{j=1}^{m} A(Y_j); \quad \bar{B}_m := \frac{1}{m}\sum_{j=1}^{m} B(Y_j), \tag{52}$$

implying that the empirical-mean envelope $G_m := \sup_{x \in \mathscr{X}} \|R_m(x)\|$ of the relative error class satisfies

$$G_m \leq \|\bar{A}_m - J_d\| + \|\bar{B}_m\| \frac{1}{c_0}. \tag{53}$$

Conclude from (53) that the uniform law in (R-ULLN) is satisfied.

Let's next demonstrate that under further moment conditions, the condition in (Bd-Grad) is satisfied as well. Suppose there exist deterministic constants $\ell_1, \ell_2 < \infty$ such that

$$\mathbb{E}\left[\left(\frac{1}{\sigma_k^*}\right)^4 \mid \mathscr{F}_{k-1}\right] \leq \ell_1 < \infty; \quad \sigma_k^* := \min_{x \neq 0} \frac{\|\bar{A}_{M_k} x\|}{\|x\|} \tag{54}$$

and

$$\mathbb{E}\left[\|B(Y)\|^4\right] \leq \ell_2 < \infty, \tag{55}$$

(Notice that σ_k^* is the minimum singular value of the matrix $\bar{A}_{M_k}$, that is, the square root of the smallest eigenvalue of the symmetric matrix $\bar{A}_{M_k}\bar{A}_{M_k}^T$.) Suppose also that the error tolerance ϵ_k in RA is chosen so that

$$\mathbb{E}\left[\epsilon_k^4 \mid \mathscr{F}_{k-1}\right] \leq \ell_3. \tag{56}$$

From (55), we see that

$$\mathbb{E}\left[\left\|\bar{B}_{M_k}\right\|^4 \mid \mathscr{F}_{k-1}\right] \leq \ell_2. \tag{57}$$

Next, we see from (50) that X_k satisfies

$$\left\|\bar{A}_{M_k}\nabla f(X_k) + \bar{B}_{M_k}\right\| \leq \epsilon_k, \tag{58}$$

implying that

$$\|\nabla f(X_k)\| \leq \|\bar{A}_{M_k}^{-1}\|\left(\epsilon_k + \|\bar{B}_{M_k}\|\right). \tag{59}$$

Also,

$$\|\bar{A}_{M_k}^{-1}\| := \max_{x \neq 0} \frac{\|\bar{A}_{M_k}^{-1} x\|}{\|x\|} = \left(\min_{x \neq 0} \frac{\|\bar{A}_{M_k} x\|}{\|x\|}\right)^{-1} = \frac{1}{\sigma_k^*}. \tag{60}$$

Hence

$$\sup_k \left\{\mathbb{E}\left[\|\nabla f(X_k)\|^2 \mid \mathscr{F}_{k-1}\right]\right\}$$
$$\leq \sup_k \left\{2\mathbb{E}\left[\left(\frac{1}{\sigma_k^*}\right)^2 \epsilon_k^2 \mid \mathscr{F}_{k-1}\right] + 2\mathbb{E}\left[\left(\frac{1}{\sigma_k^*}\right)^2 \|\bar{B}_{M_k}\|^2 \mid \mathscr{F}_{k-1}\right]\right\}$$
$$\leq 2\sup_k \left\{\left(\mathbb{E}\left[\left(\frac{1}{\sigma_k^*}\right)^4 \mid \mathscr{F}_{k-1}\right]\right)^{1/2}\left(\left(\mathbb{E}\left[\epsilon_k^4 \mid \mathscr{F}_{k-1}\right]\right)^{1/2} + \left(\mathbb{E}\left[\|\bar{B}_{M_k}\|^4 \mid \mathscr{F}_{k-1}\right]\right)^{1/2}\right)\right\}. \tag{61}$$

Using (54), (55) and (56) in (61), we see that the condition in (Bd-Grad) is satisfied. ∎

REMARK 5. Example 6.2 is a special case of a larger and useful class of random functions having random coefficients that are dependent on x, that is,

$$\nabla F(x, Y) = A(x, Y)\nabla f(x) + B(x, Y).$$

An analysis similar to what we have outlined in Example 6.2 is possible for this larger class but with stipulations on $\{A(x, Y), x \in \mathbb{R}^d\}$ and $\{B(x, Y), x \in \mathbb{R}^d\}$ replacing our current stipulations on $A(Y)$ and $B(Y)$.

6.3. Example (Linear and Polynomial Regression with Least Squares). Let's consider the classical regression problem of statistics where we wish to estimate the parameters $x^* \in \mathbb{R}^p$ of an assumed linear regression model

$$T(Y) = Qx^* + \epsilon(Y), \tag{62}$$

where $T \in \mathbb{R}$ is the "observed" dependent variable, $Q = (Q^{(1)}, Q^{(2)}, \ldots, Q^{(d)})$ is a $1 \times d$ random vector of covariates, x^* is a $d \times 1$ vector of true parameters to be estimated using data. Assume:

$$\mathbb{E}[\epsilon(Y)] = 0; \quad \text{Corr}(Q^{(j)}, \epsilon(Y)) = 0 \,\forall j; \quad \mathbb{E}[Q^T Q] < \infty; \quad \mathbb{E}[Q^T T] < \infty; \text{ and } \left(\mathbb{E}[Q^T Q]\right)^{-1} \text{ exists.} \tag{63}$$

The question of estimating x^* can be posed as the following unconstrained stochastic optimization problem:

$$\text{minimize: } f(x) := \frac{1}{2}\mathbb{E}\left[(T - Qx)^2\right]$$
$$\text{s.t. } x \in \mathbb{R}^d. \tag{64}$$

Using (63), it can be shown that x^* is the unique solution of the problem in (64).

Suppose now that we observe data $(Q_j, T_j), i = 1, 2, \ldots, n$, where $Q_j, T_j, j = 1, 2, \ldots, n$ are iid copies of (Q, T) and $n > d$. Let $\tilde{Q}_n$ represent the $n \times d$ matrix obtained by stacking $Q_j, j = 1, 2, \ldots, n$ and let $\tilde{T}_n$ be the $n \times 1$ matrix obtained by stacking $T_j, j = 1, 2, \ldots, n$. (The "tall" $n \times d$ matrix $\tilde{Q}_n$ obtained by stacking $Q_i, i = 1, 2, \ldots, n$ is traditionally called the "data matrix" [74].) The sample-path problem corresponding to (64) is

$$\text{minimize: } f_n(x) := \frac{1}{2}\frac{1}{n}\sum_{j=1}^{n}(T_j - Q_j x)^2$$
$$\text{s.t. } x \in \mathbb{R}^d. \tag{65}$$

Now observe that

$$\nabla f(x) = \mathbb{E}\left[Q^T Q\right] x - \mathbb{E}\left[Q^T T\right];$$
$$\nabla f_n(x) = \frac{1}{n}\sum_{j=1}^{n}\left(Q_j^T Q_j x - Q_j^T T_j\right) = \frac{1}{n}\left(\tilde{Q}_n^T \tilde{Q}_n x - \tilde{Q}_n^T \tilde{T}_n\right). \tag{66}$$

Notice also that under the further assumption that the data matrix $\tilde{Q}_n$ has rank d, we get the classical least squares estimator X_n^* for x^*:

$$X_n^* = \left(\tilde{Q}_n^T \tilde{Q}_n^{-1}\right)^{-1} \tilde{Q}_n^T \tilde{T}_n. \tag{67}$$

From (66), we see that the relative error class

$$\mathscr{R} := \{ \left(Q^T Q - \mathbb{E}\left[Q^T Q \right] \right) x - \left(Q^T T - \mathbb{E}\left[Q^T T \right] \right), x \in \mathbb{R}^d \},$$

and the relative error $R_m(x)$ for any $x \in \mathbb{R}^d$ satisfies

$$
\begin{aligned}
\| R_m(x) \| &= \frac{\left\| \bar{A}_m x - \bar{B}_m \right\|}{\| \mathbb{E}\left[Q^T Q \right] x - \mathbb{E}\left[Q^T T \right] \| + c_0} \\
&\leq \| \bar{A}_m \| \frac{\| x \|}{\| \mathbb{E}\left[Q^T Q \right] x - \mathbb{E}\left[Q^T T \right] \| + c_0} + \frac{\| \bar{B}_m \|}{c_0},
\end{aligned}
\tag{68}
$$

where

$$\bar{A}_m := \mathbb{E}\left[Q^T Q \right] - \frac{1}{m} \sum_{j=1}^{n} Q_j^T Q_j; \quad \bar{B}_m := \mathbb{E}\left[Q^T T \right] - \frac{1}{m} \sum_{j=1}^{n} Q_j^T T_j. \tag{69}$$

Due to the assumption in (63),

$$\bar{A}_m \overset{\mathrm{wp1}}{\to} 0; \quad \bar{B}_m \overset{\mathrm{wp1}}{\to} 0, \tag{70}$$

and since $\mathbb{E}[Q^T Q]$ is nonsingular, we can show after some algebra that

$$\sup_{x \in \mathbb{R}^d} \frac{\| x \|}{\| \mathbb{E}\left[Q^T Q \right] x - \mathbb{E}\left[Q^T T \right] \| + c_0} < \infty. \tag{71}$$

Using (71) and (70) in (68), we see that the empirical-mean envelope $G_m := \sup_{x \in \mathbb{R}^d} \| R_m(x) \| \overset{\mathrm{wp1}}{\to} 0$ and hence the uniform law in (R-ULLN) is satisfied assuring almost sure convergence of RA when implemented on this class.

Let's now demonstrate that under further conditions, the bounded gradients assumption in (Bd-Grad) is also satisfied, ensuring L_1 convergence of RA. Suppose that the following technical condition on the observed data holds. There exists $\epsilon > 0$ such that

$$\left| Q^{(i)} T \right| \geq \epsilon, \quad i = 1, 2, \ldots, d \tag{72}$$

and that $\epsilon_k < \epsilon$ for all $k \geq 1$. After some algebra, we can write the sample gradient ∇f_n in terms of the true gradient ∇f as follows:

$$\nabla f_n(x) = \Lambda_n(x) \nabla f(x) + 2 \left(\Lambda_n \mathbb{E}\left[Q^T T \right] - \tilde{Q}_n^T \tilde{T}_n \right), \tag{73}$$

where Λ_n is a $d \times d$ diagonal matrix given by

$$\Lambda_n(x) := \begin{bmatrix} \frac{n^{-1} \sum_{j=1}^{n} Q_j^{(1)} Q_j x}{\mathbb{E}\left[Q^{(1)} Q \right] x} & & \\ & \ddots & \\ & & \frac{n^{-1} \sum_{j=1}^{n} Q_j^{(d)} Q_j x}{\mathbb{E}\left[Q^{(d)} Q \right] x} \end{bmatrix}. \tag{74}$$

From (66) and since the solution X_k obtained after k iterations of RA satisfies

$$\| \nabla f_{M_k}(X_k) \| \leq \epsilon_k, \tag{75}$$

we see that

$$\| \nabla f(X_k) \| \leq \left\| \Lambda_{M_k}^{-1}(X_k) \right\| \left(\epsilon_k + 2 \left\| \tilde{Q}_{M_k}^T \tilde{T}_{M_k} \right\| \right) + 2 \left\| \mathbb{E}\left[Q^T T \right] \right\|, \tag{76}$$

and hence

$$\mathbb{E}\left[\|\nabla f(X_k)\|^2\right] \le 4\sqrt{2}\left(\left\|\mathbb{E}\left[\Lambda_{M_k}^{-1}(X_k)\right\|^4\right]\right)^{1/2}\left(\epsilon_k^4 + 16\mathbb{E}\left[\left\|Q^T T\right\|^4\right]\right)^{1/2} + 8\left\|\mathbb{E}\left[Q^T T\right]\right\|^2. \quad (77)$$

Also, some algebra yields that

$$\left| M_k^{-1}\sum_{j=1}^{M_k} Q_j^{(i)} Q_j X_k \right| \ge \left| M_k^{-1}\sum_{j=1}^{M_k} Q_j^{(i)} T_j \right| - \epsilon_k \text{ and } \left|\mathbb{E}\left[Q^{(i)}Q\right]X_k\right| \le \left|\mathbb{E}\left[Q^{(i)}T\right]\right| + \epsilon_k. \quad (78)$$

Using (78) in (74), we have that the i-th diagonal element $\Lambda_{M_k}^{-1}(X_k)(i,i)$ of the matrix $\Lambda_{M_k}^{-1}(X_k)$ satisfies

$$\mathbb{E}\left[\left|\Lambda^{-1}(X_k)(i,i)\right|^4\right] \le 16\left(\left|\mathbb{E}\left[TQ^{(i)}\right]\right|^4 + \epsilon_k^4\right)\mathbb{E}\left[\left(\left|M_k^{-1}\sum_{j=1}^{M_k} Q_j^{(i)} T_j\right| - \epsilon_k\right)^{-4}\right]$$
$$< \infty, \quad (79)$$

where we have used the technical condition in (72) to obtain the last inequality in (79). Use (79) in (77) to conclude that the sufficient condition (Bd-Grad) for L_2 convergence holds.∎

It is interesting that for the bounded gradients assumption (Bd-Grad) to hold, the further (somewhat stringent) technical condition in (72) is needed even though L_1 convergence does not imply almost sure convergence in general. Also, while the treatment in Example 6.3 was for linear regression, similar conclusions for polynomial regression hold in a straightforward manner upon introducing higher powers of x^* in (62).

6.4. Example (Karhounen-Loève Error Class). The Karhunen-Loève (KL) representation [25], analogous to the Fourier representation of a deterministic function, is a series representation of a stochastic process as a linear combination of orthonormal functions having random coefficients that are pairwise uncorrelated. Models of physical processes based on KL are among the most fruitful applications of the theory of stochastic processes, and have been extensively used within statistical estimation and finite-element settings [25] over the past few decades. For this reason, such models are also natural contexts where stochastic optimization using a method such as RA may be of direct interest.

As a simple illustration, suppose the error process $\{\nabla F(x,Y) - \nabla f(x), x \in \mathscr{X}\}$ is a second-order process, that is, $\mathbb{E}\left[\|\nabla F(x,Y) - \nabla f(x)\|^2\right] < \infty$ for each $x \in \mathscr{X}$. Then, $\{\nabla F(x,Y) - \nabla f(x), x \in \mathscr{X}\}$ admits a KL representation [2, 25], that is, we can write

$$\nabla F(x,Y) - \nabla f(x) = \sum_{j=1}^{\infty} \sqrt{\lambda_j}\, Z_j(Y)\,\phi_j(x), \quad x \in \mathscr{X}, \quad (80)$$

where $\phi_j : \mathscr{X} \to \mathbb{R}^d, j \ge 1$ form an orthonormal basis (eigenfunctions) that spans $L^2(\mathscr{X})$, the associated eigen values $\lambda_j, j \ge 1$ and coefficients $\{Z_j, j \ge 1\}$ satisfy

$$\mathbb{E}[Z_j] = 0; \quad \text{and} \quad \mathbb{E}[Z_i Z_j] = \delta_{ij}. \quad (81)$$

The convergence in (80) is L_2, that is,

$$e_n(x) := \nabla F(x,Y) - \nabla f(x) - \sum_{j=1}^{n} \sqrt{\lambda_j}\, Z_j(Y)\,\phi_j(x), \quad x \in \mathscr{X}$$

satisfies $\lim_{n\to\infty} \mathbb{E}\left[\|e_n(x)\|^2\right] = 0$ for each $x \in \mathscr{X}$.

Suppose also that the following three conditions hold:

(i) $\exists b < \infty$ such that $\forall j \geq 1$, the eigenfunctions ϕ_j satisfy

$$\sup_{x \in \mathscr{X}} \frac{\|\phi_j(x)\|}{\|\nabla f(x)\| + c_0} \leq b;$$

(ii) $\sum_{j=1}^{\infty} \lambda_j < \infty$; and

(iii) there exist $\delta(\epsilon) > 0$, $\sigma < \infty$ such that for each j:

$$\mathbb{E}\left[\left|\bar{Z}_{j,m}\right|^{2+\epsilon}\right] \leq \frac{\sigma}{m^{1+\delta(\epsilon)}}.$$

The assumption in (iii) holds under standard assumptions leading to the Bernstein inequality [78, pp. 33], that is, when $\{Z_j, j \geq 1\}$ are independent, sub-exponential random variables; there exist extensions to the context where $\{Z_j, j \geq 1\}$ are sub-exponential but dependent — see, for example, [20] and [78, Section 5].

Under (ii) and (iii), it can be shown that the sequence $\{\sum_{j=1}^{n} \sqrt{\lambda_j} Z_j(Y), n \geq 1\}$ is uniformly integrable and hence the equality in (80) holds pointwise (almost surely). Also, from (80), we see that for each $x \in \mathscr{X}$,

$$\|R_m(x)\| = \frac{\|\nabla f_m(x) - \nabla f(x)\|}{\|\nabla f(x)\| + c_0} = \frac{1}{\|\nabla f(x)\| + c_0} \times \left\|\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{\infty} \sqrt{\lambda_j}\, Z_j(Y_i)\, \phi_j(x)\right\|,$$

$$= \frac{1}{\|\nabla f(x)\| + c_0} \times \left\|\sum_{j=1}^{\infty} \sqrt{\lambda_j}\, \bar{Z}_{j,m}\, \phi_j(x)\right\|$$

$$\leq \frac{b}{c_0} \sum_{j=1}^{\infty} \sqrt{\lambda_j}\, \left|\bar{Z}_{j,m}\right|, \tag{82}$$

where the last inequality in (82) follows from (i) above. Now notice that

$$\sum_{m=1}^{\infty} P\left(\sum_{j=1}^{\infty} \sqrt{\lambda_j}\, |\bar{Z}_{j,m}| > \nu\right) \leq \sum_{m=1}^{\infty} \frac{1}{\nu^{1+\epsilon}} \mathbb{E}\left[\left(\sum_{j=1}^{\infty} \sqrt{\lambda_j}\, |\bar{Z}_{j,m}|\right)^{1+\epsilon}\right]$$

$$\leq \frac{1}{\nu^{1+\epsilon}} \left(\sum_{j=1}^{\infty} \sqrt{\lambda_j}\right)^{\epsilon} \sum_{m=1}^{\infty}\sum_{j=1}^{\infty} \mathbb{E}\left[\sqrt{\lambda_j}\, |\bar{Z}_{j,m}|^{1+\epsilon}\right]$$

$$\leq \frac{1}{\nu^{1+\epsilon}} \left(\sum_{j=1}^{\infty} \lambda_j\right)^{\frac{1+\epsilon}{2}} \left(\sum_{m=1}^{\infty} \mathbb{E}\left[\left|\bar{Z}_{j,m}\right|^{2+2\epsilon}\right]\right)^{1/2}$$

$$< \infty, \tag{83}$$

where the last inequality in (83) follows from (ii) and (iii) above. We now conclude from (83) and Borel-Cantelli's first lemma [7, pp. 59,60] that

$$\sum_{j=1}^{\infty} \sqrt{\lambda_j} |\bar{Z}_{j,m}| \xrightarrow{\text{wp1}} 0. \tag{84}$$

Using (84) in (82), we see that

$$\sup_{x \in \mathscr{X}} \|R_m(x)\| \xrightarrow{\text{wp1}} 0, \tag{85}$$

implying that Theorem 1 guarantees strong consistency if RA is implemented on this class with $M_k \to \infty$ and $\epsilon_k \to 0$.

Using arguments similar to those in (45)–(47) of Example I, we can also conclude that the key condition in (Bd-Grad) holds. ∎

7. ORACLE COMPLEXITY We now provide results that quantify the non-asymptotic convergence rates (also known as *complexity*) of RA when expressed in terms of the total number of oracle calls expended. Recall that unlike SGD, RA algorithms may expend a varying number of oracle calls during each outer iteration, that is, between subsample updates. Specifically, the total number of oracle calls expended by RA after k iterations is:

$$W_k := \sum_{j=1}^{k} N_j \times M_j, \tag{86}$$

where N_j is the number of solver steps (or inner iterations), and M_j is the adapted sample-size during the j-th iteration. Notice that the expression for the total work appearing in (86) assumes a "linear sum" structure; such structure is not always valid but holds for many common optimization solvers such as gradient descent [50, Chapter 2], accelerated gradient descent [50, Chapter 2], momentum method [13, Section 7], and line search with L-BFGS [51], since these methods require only gradient evaluations at each inner iteration.

The oracle complexity attained by RA depends on the speed of the solver employed to execute the inner iterations. Accordingly, we present our oracle complexity results as a function of whether the solver used in the inner iterations is *linearly convergent* or *sub-linearly convergent.* — see Definition 4. Theorem 5 that follows is intended to address problems having strongly convex sample-path problems while Theorem 6 addresses problems having (merely) convex or non-convex sample-path problems.

THEOREM 5 (**Expected Oracle Complexity for Linear Solvers**). *Suppose the Method* $\mathfrak{M}$ *within RA exhibits* $(\rho_{M_k}^{\mathfrak{M}}, C_{M_k}^{\mathfrak{M}})$-*linear convergence (see Definition 4) when executed on the sample path problem* (S_{M_k}) *starting at initial point* X_{k-1}, *where* $0 < \rho_{M_k}^{\mathfrak{M}} \leq \rho_{\mathfrak{M}} < 1$ *and* $C_{M_k}^{\mathfrak{M}} \leq C_{\mathfrak{M}}$, *and* $\rho_{\mathfrak{M}} < 1, C_{\mathfrak{M}} < \infty$ *are some constants. Let the sample size sequence* $\{M_k, k \geq 1\}$ *and the error tolerance sequence* $\{\epsilon_k, k \geq 1\}$ *be chosen so that*

$$M_k := C_{1,k} M_{k-1}, \qquad\qquad C_{1,k} = c_1, 1 < c_1 < \infty; \tag{87}$$

$$\epsilon_k := \frac{C_{2,k}}{\sqrt{M_k}}, \qquad\qquad C_{2,k} \in [\underline{c}_2, c_2], 0 < \underline{c}_2 \leq c_2 < \infty, \tag{88}$$

$$\text{(C.3)}$$

where $C_{1,k}, C_{2,k} \in \mathscr{F}_{k-1}$, *and* $c_1, \bar{c}_1, c_2, \underline{c}_2$ *are (non-random) constants.*
If the postulates of Theorem 3 hold, then, for each $k \geq 1$,

$$\mathbb{E}[W_k] = O\left(\frac{A^2 m_1}{(c_1 - 1)\epsilon^2 \log(1/\rho_{\mathfrak{M}})}\right) \implies \mathbb{E}[\nabla f(X_k)] \leq \epsilon, \tag{89}$$

with

$$A = \frac{\sqrt{c_1}(c_2 + c_0\sigma + \sigma\ell_{2,\infty})}{\sqrt{m_1}}. \tag{90}$$

If the postulates of Theorem 4 hold and (CLT-sc) *is satisfied, then* (89) *holds with*

$$A = \sqrt{c_1}\left(\frac{2(c_2 + c_0\sigma_0/\delta_0)}{\sqrt{m_1}}\right) + \left(\frac{\kappa_2}{\sqrt{m_1}}\right)^{1/\alpha}. \tag{91}$$

Proof. From the definition of linear convergence (Definition 4), we see that for any

$$N_k \geq \frac{\log\left(C_{\mathfrak{M}} \|\nabla f_{M_k}(X_{k-1})\| / \epsilon_k\right)}{\log(1/\rho_{\mathfrak{M}})}$$

the termination condition $\|\nabla f_{M_k}(X_k)\| \leq \epsilon_k$ is satisfied. We can choose N_k to be the smallest number such that the condition $\|\nabla f_{M_k}(X_k)\| \leq \epsilon_k$ is satisfied. Therefore,

$$N_k = \left\lceil \frac{\log\left(C_{\mathfrak{M}}\|\nabla f_{M_k}(X_{k-1})\|/\epsilon_k\right)}{\log(1/\rho_{\mathfrak{M}})} \right\rceil \leq 1 + \frac{\log\left(C_{\mathfrak{M}}\|\nabla f_{M_k}(X_{k-1})\|/\epsilon_k\right)}{\log(1/\rho_{\mathfrak{M}})}$$

Taking expectation, using (12), and applying Jensen's inequality on a concave function, we have,

$$\mathbb{E}[N_k] \leq 1 + \frac{\log(C_{\mathfrak{M}}) + \log(\mathbb{E}[\|\nabla f_{M_k}(X_{k-1})\|/\epsilon_k])}{\log(1/\rho_{\mathfrak{M}})}$$
$$\leq 1 + \frac{\log(C_{\mathfrak{M}}) + \log(\mathbb{E}[\|\nabla f_{M_k}(X_{k-1}) - \nabla f(X_{k-1})\| + \|\nabla f(X_{k-1})\|/\epsilon_k])}{\log(1/\rho_{\mathfrak{M}})}$$
$$\leq 1 + \frac{\log(C_{\mathfrak{M}}) + \log(\mathbb{E}[\|G_{M_k}\|(\|\nabla f(X_{k-1})\| + c_0) + \|\nabla f(X_{k-1})\|/\epsilon_k])}{\log(1/\rho_{\mathfrak{M}})}$$
$$\leq 1 + \frac{\log(C_{\mathfrak{M}})}{\log(1/\rho_{\mathfrak{M}})} + \frac{\log\left(\sigma\mathbb{E}[(\|\nabla f(X_{k-1})\| + c_0)] + \sqrt{M_k}\mathbb{E}[\|\nabla f(X_{k-1})\|]\right) - \log \underline{c_2}}{\log(1/\rho_{\mathfrak{M}})}$$

where the last inequality is due to the assumption (CLT-sc) and (88). Now from Theorem 3 and 4, we have that

$$\mathbb{E}[\|\nabla f(X_{k-1})\|] \leq C\left(\frac{1}{\sqrt{c_1}}\right)^{k-1}.$$

Substituting this in the above inequality, we get,

$$\mathbb{E}[N_k] \leq 1 + \frac{\log(C_{\mathfrak{M}})}{\log(1/\rho_{\mathfrak{M}})} + \frac{\log\left(\sigma(Cc_1^{-(k-1/2)} + c_0) + \sqrt{m_1}C\right) - \log \underline{c_2}}{\log(1/\rho_{\mathfrak{M}})},$$
$$=: \frac{C_u}{\log(1/\rho_{\mathfrak{M}})}$$

where

$$C_u = \log\left(\frac{C_{\mathfrak{M}}\left(\sigma(Cc_1^{-(k-1/2)} + c_0) + \sqrt{m_1}C\right)}{\rho_{\mathfrak{M}}\underline{c_2}}\right).$$

The expected work complexity to achieve $\mathbb{E}[\|\nabla f(X_k)\|] \leq \epsilon$ is then

$$\mathbb{E}[W_{K(\epsilon)}] = \sum_{j=1}^{K(\epsilon)} M_j \mathbb{E}[N_j] \leq \sum_{j=1}^{K(\epsilon)} m_1 \bar{c}_1^{j-1} \frac{C_u}{\log(1/\rho_{\mathfrak{M}})}$$
$$\leq m_1 \frac{C_u}{\log(1/\rho_{\mathfrak{M}})} \frac{c_1^{K(\epsilon)} - 1}{c_1 - 1}. \tag{92}$$

Corollaries 1 and 2, imply $K(\epsilon) = \frac{\log\left(\frac{A}{\epsilon}\right)}{\log \sqrt{c_1}}$, and $c_1^{K(\epsilon)} = \left(\frac{A}{\epsilon}\right)^2$. Plugging in (92), we get

$$\mathbb{E}[W_{K(\epsilon)}] \leq \frac{m_1}{c_1 - 1} \frac{C_u}{\log(1/\rho_{\mathfrak{M}})} c_1^{K(\epsilon)}$$
$$\leq \frac{m_1}{c_1 - 1} \frac{C_u}{\log(1/\rho_{\mathfrak{M}})} \left(\frac{A}{\epsilon}\right)^2. \blacksquare$$

Observe that Theorem 5 signals the optimal $O(\epsilon^{-2})$ oracle complexity, with the constant A involving the rate of increase of sample size c_1, the inherent "variability" σ of the stochastic oracle, and the extent of the excursions $\ell_{2,\infty}$ of RA's iterates. Theorem 5 assumes that $C_{1,k} = c_1$, that is, the sample sizes are increased at a fixed rate c_1; it can be shown that allowing the rate of increase of the sample sizes to be in an interval introduces a slight inflation in the expected oracle complexity rate.

Next, we prove a result analogous to Theorem 5 but when a sub-linear solver is used when solving the sample-path problems.

THEOREM 6 (**Expected Oracle Complexity for Sublinear Solvers**). *Suppose the Method $\mathfrak{M}$ within RA exhibits $(1/2, C_{M_k}^{\mathfrak{M}})$-sublinear convergence (see Definition 4) when executed on the sample path problem (S_{M_k}) starting at initial point X_{k-1}, where $0 < C_{M_k}^{\mathfrak{M}} \leq C_{\mathfrak{M}}$, and $C_{\mathfrak{M}} < \infty$ is some constant. Let the sample size sequence$\{M_k, k \geq 1\}$ and the error tolerance sequence $\epsilon_k, k \geq 1\}$ be chosen according to (87), and suppose further that there exists $C_A < \infty$ such that*

$$\sup_k \mathbb{E}[f(X_k)] < \frac{C_A}{2}; \quad \mathbb{E}[\inf_{x \in \mathbb{R}^d} F(x, Y)] > -\frac{C_A}{2}. \tag{Bd-Fn}$$

If the postulates of Theorem 3 or Theorem 4 hold and (CLT-sc) is satisfied, then, for each $k \geq 1$,

$$\mathbb{E}[W_k] = O\left(\frac{A^4 m_1 C_{\mathfrak{M}} C_A}{(c_1^2 - 1)\underline{c_2}^2 \epsilon^4}\right) \implies \mathbb{E}[\nabla f(X_k)] \leq \epsilon, \tag{93}$$

where A is as in (90) or (91) (according to whether Theorem 3 or Theorem 4 hold), and

$$C_u = \frac{C_{\mathfrak{M}}\left(\sigma(C c_1^{-(k-1/2)} + c_0) + \sqrt{m_1}C\right)}{\underline{c_2}}. \tag{94}$$

Proof. From the definition of sublinear convergence (Definition 4), we see that for any

$$N_k \geq \frac{C_{\mathfrak{M}}(f_{M_k}(X_{k-1}) - f_{M_k}(X_k))}{\epsilon_k^2}$$

the termination condition $\|\nabla f_{M_k}(X_k)\| \leq \epsilon_k$ is satisfied. We can choose N_k to be the smallest number such that $\|\nabla f_{M_k}(X_k)\| \leq \epsilon_k$ condition is satisfied. Therefore,

$$N_k = \left\lceil \frac{C_{\mathfrak{M}}(f_{M_k}(X_{k-1}) - f_{M_k}(X_k))}{\epsilon_k^2} \right\rceil \leq \frac{2C_{\mathfrak{M}}(f_{M_k}(X_{k-1}) - f_{M_k}(X_k))}{\epsilon_k^2}$$

Taking expectation, we have,

$$\mathbb{E}[N_k] \leq \mathbb{E}\left[\frac{2C_{\mathfrak{M}}(f_{M_k}(X_{k-1}) - f_{M_k}(X_k))}{\epsilon_k}\right]$$
$$\leq \frac{2M_k C_{\mathfrak{M}}(\mathbb{E}[f_{M_k}(X_{k-1}) - f_{M_k}(X_k)])}{\underline{c_2}^2}$$

Now using (Bd-Fn), we can show

$$\mathbb{E}[f_{M_k}(X_{k-1}) - f_{M_k}(X_k)] \leq \mathbb{E}[f(X_{k-1})] - \mathbb{E}[\inf_{x \in \mathbb{R}^d} f_{M_k}(x)]$$
$$\leq C_A.$$

Therefore, we have,

$$\mathbb{E}[N_k] \leq \frac{2M_k C_{\mathfrak{M}} C_A}{\underline{c_2}^2}$$

Now from Theorem 3 and 4, we have that

$$\mathbb{E}[\|\nabla f(X_{k-1})\|] \leq C \left(\frac{1}{\sqrt{c_1}}\right)^{k-1}.$$

Following the same steps employed in Theorem 5, we obtain

$$\mathbb{E}[W_{K(\epsilon)}] = \sum_{j=1}^{K(\epsilon)} M_j \mathbb{E}[N_j] \leq \sum_{j=1}^{K(\epsilon)} m_1^2 c_1^{2(j-1)} \frac{2C_{\mathfrak{m}} C_A}{\underline{c_2}^2}$$
$$\leq \frac{2m_1^2 C_{\mathfrak{m}} C_A}{\underline{c_2}^2} \frac{c_1^{2K(\epsilon)} - 1}{c_1^2 - 1}.$$

From Corollaries 1 and 2, we have $K(\epsilon) = \frac{\log\left(\frac{A}{\epsilon}\right)}{\log \sqrt{c_1}}$ and substituting this into the above inequality gives the desired result. ∎

   The derived complexities in Theorem 5 for linear solvers and Theorem 6 for sublinear solvers should be loosely interpreted as $O(\epsilon^{-2})$ and $O(\epsilon^{-4})$, respectively. The former complexity matches the best achievable rate for stochastic optimization on smooth strongly convex functions, while the latter rate matches the $O(\epsilon^{-4})$ lower bound for smooth non-convex stochastic optimization [3].

   Also, Theorem 6 relies on the technical condition in (Bd-Fn). To understand this condition, notice that we have assumed little about the nature of the solutions obtained by the solvers used within RA on successive iterations. So, presumably, since f can be non-convex, the solver in use and the problem landscape might collude in a way that RA's trajectory oscillates between first-order critical point estimators in regions with a high objective and those with a low objective. The condition in (Bd-Fn) prevents such oscillation from being extreme.

   Recall that the k-th (outer) iteration of RA is terminated when the norm $\|\nabla f_{M_k}(X_k)\|$ of the subsampled gradient at the incumbent iterate falls below the threshold ϵ_k, that is, if $X_{k-1,t}$ is the t-th iterate obtained from the employed solver during the k-th iteration, then the number of inner steps

$$N_k := \min\left\{t : \|\nabla f_{M_k}(X_{k,t})\| \leq \epsilon_k\right\}.$$

Theorem 5 and Theorem 6 consider the context where the tolerance ϵ_k is chosen as $\epsilon_k = C_{2,k}/\sqrt{M}_k$, implying that

$$N_k := \min\{t : \|\nabla f_{M_K}(X_{k,t})\| \leq \frac{C_{2,k}}{\sqrt{M_k}}\}. \tag{95}$$

   An alternative to the stopping criterion in (95), and one that is consistent with practice within deterministic solvers, is to terminate the inner iterations of the k-th outer iteration when the relative error drops below a threshold, that is,

$$N_k := \min\left\{t : \|\nabla f_{M_k}(X_{k,t})\| \leq \Gamma_k \|\nabla f_{M_k}(X_{k-1})\|\right\}, \tag{96}$$

where $\Gamma_k \in \mathscr{F}_{k-1}$.

   To characterize the expected oracle complexity rate when the stopping criterion is chosen according to the "practical condition" in (96), we need the following alternate model of progress for sublinear solvers. A solver is said to exhibit *C-sublinear convergence* $(C < \infty)$ if the gradient norm at iterate x_n obtained after executing n steps of the algorithm satisfies

$$\|\nabla f(x_n)\| \leq C \|\nabla f(x_0)\| n^{-1/2}. \tag{97}$$

The model of progress in (97) is possibly more typical of behavior in practice than that in (4) but there appears to be no straightforward justification such as in (5). Using the alternate notion of sublinear convergence in (97), we can make the following remark about the oracle complexity when using the practical termination condition in (96).

REMARK 6. Suppose we set

$$\epsilon_k := \Gamma_k \|\nabla f_{M_k}(X_{k-1})\|,$$

where $\Gamma_k \geq \gamma > 0, \Gamma_k \in \mathscr{F}_{k-1}$. Let the postulates of Theorem 3 or Theorem 4 hold and let (CLT-sc) be satisfied. If a linear solver or a sublinear solver satisfying (97) is used within RA, then, for each $k \geq 1$,

$$\mathbb{E}[W_k] \geq O(\epsilon^{-2}) \implies \mathbb{E}[\nabla f(X_k)] \leq \epsilon.$$

Since the above statement can be proved using the same steps as in Theorems 3, 4 and 5 we do not include a proof.

8. CENTRAL LIMIT THEOREM (CLT) We next treat the question of whether the sequence $\{\|\nabla f(X_k)\|, k \geq 1\}$ exhibits a CLT. Unlike the complexity results from Section 5 and Section 7 which are non-asymptotic in the sense that they are valid for all $k \geq 1$, the CLT we present here holds only asymptotically for $k \gg 1$. CLTs say less than the complexity results about the behavior of $\{\|\nabla f(X_k)\|, k \geq 1\}$ for small k, but say more for large k in the sense that they characterize the exact asymptotic distribution of $\|\nabla f(X_k)\|$. In particular, they precisely quantify the effect of the curvature of the objective function f at the critical points on the rate of convergence. For this reason, CLTs are crucial building blocks for constructing confidence regions especially in the service of terminating an algorithm.

Before we state the main result, we state an auxiliary result (without proof) that demonstrates the regular behavior of the scaled empirical error $\sqrt{M_k}(\nabla f_{M_k}(X_k) - \nabla f_{M_k}(X^*))$, where the random variable X^* is such that $X_k \overset{\text{wp1}}{\to} X^* \in \mathscr{X}^*$. A proof of Lemma 1 follows along lines similar to that of [75, Theorem 6.3].

LEMMA 1. *Suppose there exists a non-random non-increasing function $H : [0,1] \to \mathbb{R}^+$ such that*

$$\lim_{A \to \infty} \limsup_{n \to \infty} P\left(\sup_{u>0} \frac{H_1(u, \mathscr{R}, P_n)}{H(u)} > A\right) = 0; \quad \int_0^1 H^{1/2}(u) \, du < \infty. \tag{98}$$

Define the δ-ball around the set $\mathscr{X}^*$ of first-order critical points :

$$\mathscr{X}^*(\delta) := \bigcup_{x^* \in \mathscr{X}^*} \{x \in \mathscr{X} : \mathbb{E}\left[\|R(x,Y) - R(x^*,Y)\|\right] \leq \delta\}, \quad \delta > 0.$$

(a) *There exists $\delta > 0$ such that for any $x^* \in \mathscr{X}^*$,*

$$\lim_{n \to \infty} P\left(\sup_{x \in \mathscr{X}^*(\delta)} \sqrt{n} \|\nabla f_n(x) - \nabla f_n(x^*)\| > \eta\right) < \eta. \tag{99}$$

(b) *If the sequence of $\{X_k, k \geq 1\}$ of RA's iterates is such that $X_k \overset{\text{wp1}}{\to} X^*$ for some well-defined random variable X^*, then*

$$\sqrt{M_k} \|\nabla f_{M_k}(X_k) - \nabla f_{M_k}(X^*)\| \overset{\text{d}}{\to} 0. \tag{100}$$

We are now ready to state the main result of this section. In the result, we assume that the postulates of Theorem 1 hold, implying that $\|\nabla f(X_k)\| \overset{\text{wp1}}{\to} 0$. To avoid contexts where $\|\nabla f(X_k)\| \overset{\text{wp1}}{\to} 0$ but the sequence $\{X_k, k \geq 1\}$ does not converge to anything, we also assume that there exists a random variable X^* such that $X_k \overset{\text{wp1}}{\to} X^*$.

THEOREM 7 **(Central Limit Theorem)**. *Let the postulates of Theorem 1(b) hold along with the condition*

$$\sqrt{M_k}\,\epsilon_k \overset{d}{\to} 0. \tag{C.2(b)}$$

Suppose also that $X_k \overset{wp1}{\to} X^* \in \mathscr{X}^*$ as $k \to \infty$. Denoting the covariance matrix

$$\Sigma_x := \int \nabla F(x,y)\nabla F(x,y)^T\,P(dy), \quad x \in \mathscr{X}^*.$$

Then we have as $k \to \infty$,

$$\sqrt{M_k}\,\nabla f(X_k) \overset{d}{\to} N\left(0, \Sigma_{X^*}\right). \tag{101}$$

If we further suppose that the objective function $f : \mathbb{R}^d \to \mathbb{R}$ is twice differentiable in $\mathscr{X}^*$ with invertible second-derivative matrix, and denoting

$$V(X^*) := \nabla^2 f(X^*)^{-1}\,\Sigma_{X^*}\,(\nabla^2 f(X^*)^{-1})^T.$$

Then we have that as $k \to \infty$,

$$\sqrt{M_k}(X_k - X^*) \overset{d}{\to} N(0, V(X^*)). \tag{102}$$

Proof. Observe that

$$\begin{aligned}
\sqrt{M_k}\nabla f(X_k) &= \sqrt{M_k}\left(\nabla f(X_k) - \nabla f_{M_k}(X_k)\right) + \sqrt{M_k}\nabla f_{M_k}(X_k) \\
&= \sqrt{M_k}\int \nabla F(X_k, Y)\,d(P - P_{M_k}) + \sqrt{M_k}\nabla f_{M_k}(X_k) \\
&= \sqrt{M_k}\int \left(\nabla F(X_k, Y) - \nabla F(X^*, Y)\right)d(P - P_{M_k}) \\
&\qquad + \sqrt{M_k}\int \nabla F(X^*, Y)\,d(P - P_{M_k}) + \sqrt{M_k}\nabla f_{M_k}(X_k).
\end{aligned} \tag{103}$$

Lemma 1 guarantees that

$$\sqrt{M_k}\int \left(\nabla F(X_k, Y) - \nabla F(X^*, Y)\right)d(P - P_{M_k}) \overset{p}{\to} 0, \tag{104}$$

and condition C.2(a) ensures that

$$\sqrt{M_k}\nabla f_{M_k}(X_k) \overset{d}{\to} 0, \tag{105}$$

We now plug (104) and (105) in (103) to get

$$\sqrt{M_k}\nabla f(X_k) \overset{d}{\to} N\left(0, \Sigma(X^*)\right), \tag{106}$$

and the first assertion of the theorem (appearing in (101)) holds.

To prove the second assertion (appearing in (102)), we see that since f is twice differentiable on $\mathscr{X}^*$ with a non-singular second-derivative $\nabla^2 f$, there exists a neighborhood $B(X^*, \epsilon)$ where

$$\nabla f(X_k) = \nabla f(X_k) - \nabla f(X^*) = \nabla^2 f(X^*)(X_k - X^*) + o_p(\|X_k - X^*\|). \tag{107}$$

On $B(X^*, \epsilon)$, we can then can write

$$\sqrt{M_k}\nabla f(X_k) = \sqrt{M_k}\nabla^2 f(X^*)(X_k - X^*) + o_p(\sqrt{M_k}\|X_k - X^*\|). \tag{108}$$

Now use (106) in (108) to see that (102) holds.■

The reader will observe that the random CLT in Theorem 7 is different in nature than Theorem 12 in [35] (also see [70]) which is on the set of global minimizers, with the implication that its application to constructing confidence sets entails the practically difficult task of obtaining a global minimum estimator. Theorem 7 is also different from the classical CLT in stochastic approximation, e.g., Theorem 1 in [57], which in effect assumes strong convexity. Theorem 7 should instead be seen as being analogous to Theorem 5.2 in [55] (see also [39, Ch. 10]).

9. Numerical Experiments   To investigate the empirical performance of RA, we ran a setting of RA alongside Adam or SGD (two highly popular stochastic gradient based methods) in three different contexts: least-squares, logistic regression, and a convolutional neural net (CNN), where the objective functions are strongly convex, convex, and non-convex, respectively.

All algorithms were implemented in Python. The Tensorflow library was used for constructing the models and for computing gradients. Five different random runs were performed for each algorithm in each experimental setting. The same random seed initialization was set at the first run for each algorithm.

To compare performance of the algorithms, we measure loss as a function of the total work performed by each algorithm. We display this in a single plot for each experiment: The y-axis of each plot displays $\log(f(x) - f^*)$, where $f(x)$ is the true loss function evaluated at x, and f^* is the "optimal" loss value achieved by running a deterministic solver on the true loss for many iterations until convergence ($\|\nabla f(x)\|_2 \leq 10^{-8}$). The solid lines indicate the median loss, and the shaded region indicates the interquartile range of the losses from the five repeated runs. The loss is displayed as a function of the total number of individual gradients evaluated on the x-axis (referred to as "Total Work"). For example, if RA takes 5 inner iterations with a batch size of 100, then this is measured as 500 individual gradient evaluations. This is done so that RA and SGD/Adam are compared on similar ground, as RA is typically much more computationally intensive on a per-iteration basis.

9.1. Models and Datasets   For the least-squares experiment, we generated a dataset of 60,000 samples with 500 predictor variables. The response variable was set to be a linear function of the explanatory variables with i.i.d. standard Gaussian noise. The condition number of the covariance matrix (Hessian of the true objective function) was approximately 10^5. Mean squared error of the residual error terms was used as the objective function.

The logistic regression experiment used the fashion MNIST dataset (consisting of 60,000 28×28 greyscale images with 10 possible classes), and the CNN experiment used the LeNet CNN with the Cifar-10 dataset (60,000 32×32 RBG images with 10 possible classes). For both experiments, cross-entropy loss was used with an l_2 penalty term of $\lambda = 1/n$, where n is the number of observations in each dataset.

9.2. Algorithms   The implementation of RA was identical across each experiment. We used a sample increase rate of $q := 1.05$, with a starting batch size of $M_0 := 32$. L-BFGS with backtracking line search was used as the deterministic solver for the inner iterations. Notably, the stored gradient and iterate differences used in L-BFGS were carried across outer iterations instead of restarting L-BFGS for every new sample-path problem.

The threshold ϵ_k for terminating L-BFGS during the kth iteration was set to be $\hat{\sigma}_k^2/\sqrt{M_k}$, where $\hat{\sigma}_k^2$ is defined to be the trace of the sample gradient covariance matrix of $\nabla F_{S_\sigma}(X_k)$ for some subset of RA's current sample of size $S_\sigma := \min\{128, M_k\}$.

For implementations of Adam and SGD, the best two step sizes (in terms of final loss) were chosen from the set of $\{10^j\}, j \in \mathbb{Z}$. Step sizes were held constant throughout each run. For Adam's additional two hyper-parameters, the default Tensorflow values were used. Batch sizes for SGD and Adam were set at the default 32.

9.3. Least-Squares   Results for the least squares experiment displayed in Figure 1. The optimally-tuned SGD (step size of 10^{-5}) performs similarly to RA, with both achieving comparable loss value after 40 epochs. The loss obtained by the second best step size for SGD is substantially higher than that of RA and the best tuned SGD.
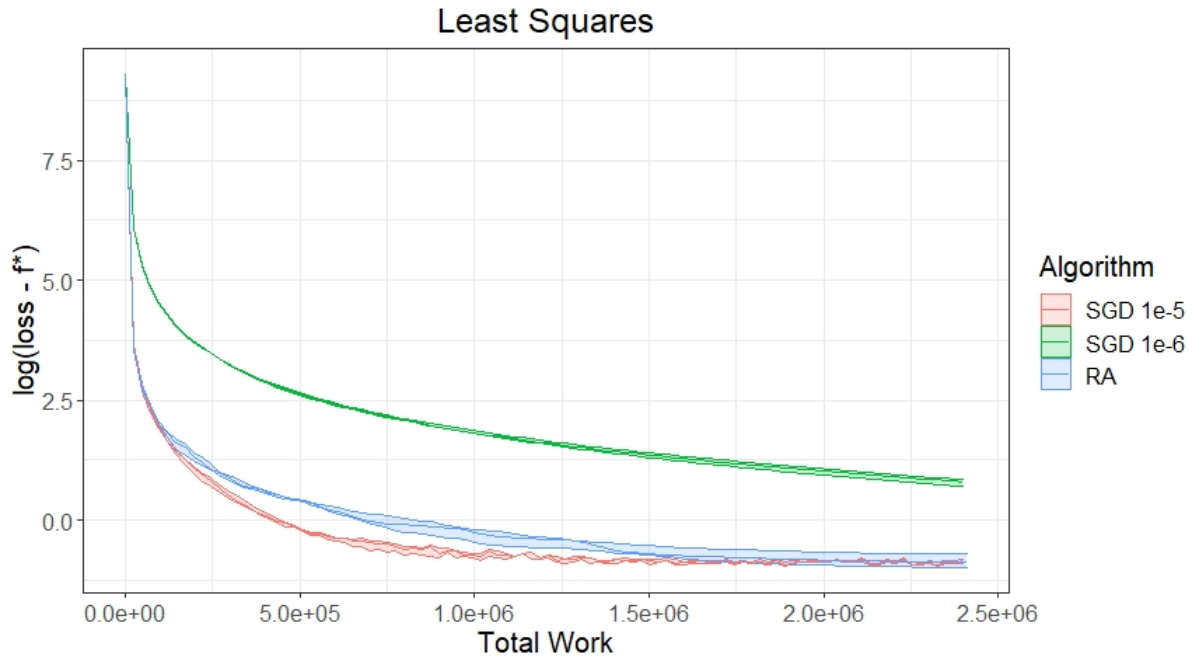
FIGURE 1. **Least Squares**

9.4. Multiple Logistic Regression   Results for the logistic regression experiment are displayed in Figure 2. The optimally-tuned SGD (step size of 10^{-2}) achieves similar loss to RA, but has considerably more variability with respect to the loss values. Using a smaller step size with SGD smooths out the algorithm's loss trajectory, but the descent is much slower as a result.
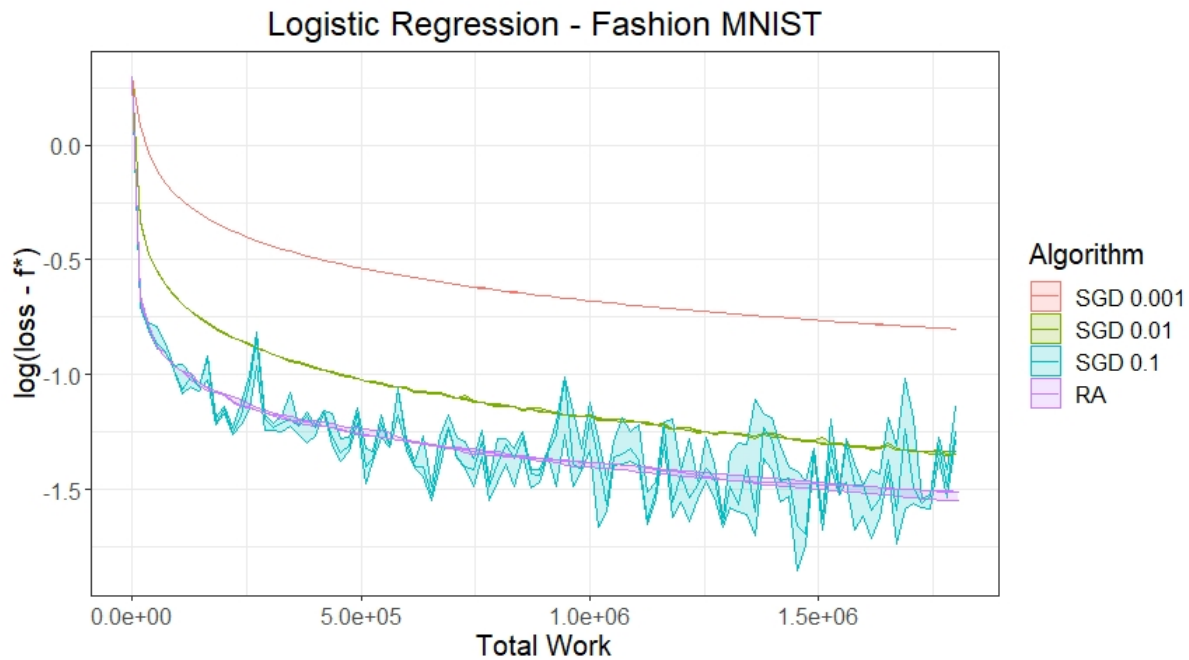

FIGURE 2. **Logistic Regression on Fashion MNIST**

9.5. LeNet on Cifar-10   Results for the CNN experiment are displayed in Figure 3. The optimally-tuned Adam (step size of 10^{-4}) achieves comparable loss to RA, but has slightly more variability with respect to the loss values. Step size parameters of 10^{-3} and 10^{-5} show deteriorating performance (in terms of variability and loss, respectively) for Adam.
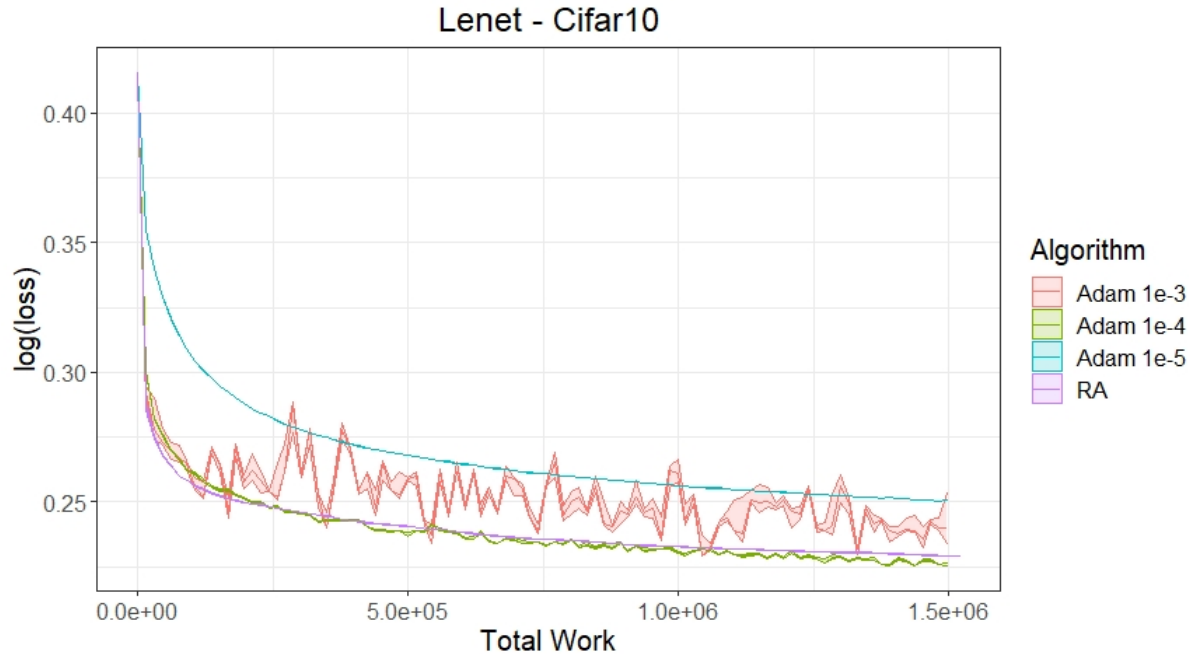

FIGURE 3. **LeNet on Cifar-10**

9.6. Discussion of Numerical Results

9.6.1. Hyper-parameter Tuning   The results from the numerical experiments illustrate some distinct features of RA compared to Adam or SGD. One clear advantage of RA is the lack of need for hyper-parameter tuning, due to its incorporation of a deterministic solver on the sample-path problems. Adam's and SGD's performance, on the other hand, is highly dependent upon the user selecting an appropriate step size (in addition to any other parameters). This phenomenon is demonstrated across all experiments, where only one choice of step size in SGD/Adam generates a loss trajectory comparable to RA. Ideally, solving an optimization problem would not require the user to have to hand-tune parameters for every individual problem. Thus, it may be argued that RA dispenses with tuning in a stochastic optimization context, demonstrating robust performance across several settings.

9.6.2. Asymptotic Consistency   From a theoretical perspective, RA guarantees several types of convergence under fairly standard assumptions on the objective function and oracle. On the other hand, SGD and Adam converge (in expectation) only if their step-sizes approach 0 at an appropriate rate. However, this often results in much poorer finite-time performance, which is likely why Tensorflow's and PyTorch's implementations of SGD and Adam use constant step-sizes by default (although more advanced techniques, such as cycling the step-size or reducing it adaptively, are available). RA seems to bridge this gap by provably being able to converge while at the same time showing robust finite-time performance without any need for parameter tuning.

References

[1] Agarwal A, Bartlett PL, Ravikumar P, Wainwright MJ (2012) Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. *IEEE transactions on information theory* 58(5):3235–3249, ISSN 0018-9448.

[2] Alexanderian A (2015) A brief note on the karhunen-loève expansion.

[3] Arjevani Y, Carmon Y, Duchi JC, Foster DJ, Srebro N, Woodworth B (2019) Lower bounds for non-convex stochastic optimization. *arXiv preprint arXiv:1912.02365* .

[4] Bartlett PL, Jordan MI, McAuliffe JD (2006) Convexity, classification, and risk bounds. *Journal of the American Statistical Association* 101(473):138–156, ISSN 0162-1459.

[5] Berahas AS, Nocedal J, Takác M (2016) A multi-batch L-BFGS method for machine learning. *Advances in Neural Information Processing Systems 29*, 1055–1063 (Curran Associates, Inc.), URL http://papers.nips.cc/paper/6145-a-multi-batch-l-bfgs-method-for-machine-learning.pdf.

[6] Bertsekas DP (2019) *Reinforcement Learning and Optimal Control* (Athena Scientific).

[7] Billingsley P (1995) *Probability and Measure* (New York, NY: Wiley).

[8] Blanchet J, Cartis C, Menickelly M, Scheinberg K (2019) Convergence rate analysis of a stochastic trust-region method via supermartingales. *INFORMS Journal on Optimization* 1(2):92–119, URL http://dx.doi.org/10.1287/ijoo.2019.0016.

[9] Bollapragada R, Byrd R, Nocedal J (2018) Adaptive sampling strategies for stochastic optimization. *SIAM Journal on Optimization* 28(4):3312–3343, URL http://dx.doi.org/10.1137/17m1154679.

[10] Bollapragada R, Byrd RH, Nocedal J (2018) Exact and inexact subsampled Newton methods for optimization. *IMA Journal of Numerical Analysis* 39(2):545–578, URL http://dx.doi.org/10.1093/imanum/dry009.

[11] Bollapragada R, Nocedal J, Mudigere D, Shi HJ, Tang PTP (2018) A progressive batching L-BFGS method for machine learning. Dy J, Krause A, eds., *Proceedings of the 35th International Conference on Machine Learning*, volume 80, 620–629 (PMLR), URL http://proceedings.mlr.press/v80/bollapragada18a.html.

[12] Bottou L, Bousquet O (2011) The tradeoffs of large-scale learning. *Optimization for Machine Learning*, 351– (The MIT Press), ISBN 026201646X.

[13] Bottou L, Curtis F, Nocedal J (2018) Optimization methods for large-scale machine learning .

[14] Bottou L, Curtis FE, Nocedal J (2018) Optimization methods for large-scale machine learning. *Siam Review* 60(2):223–311.

[15] Byrd RH, Chin GM, Nocedal J, Wu Y (2012) Sample size selection in optimization methods for machine learning. *Mathematical Programming* 134(1):127–155, URL http://dx.doi.org/10.1007/s10107-012-0572-5.

[16] Cartis C, Scheinberg K (2018) Global convergence rate analysis of unconstrained optimization methods based on probabilistic models. *Mathematical Programming* 169(2):337–375.

[17] Chen H, Schmeiser BW (2001) Stochastic root finding via retrospective approximation. *IIE Transactions* 33:259–275.

[18] Curtis FE, Scheinberg K (2017) Optimization methods for supervised machine learning: From linear models to deep learning. *Leading Developments from INFORMS Communities*, 89–114 (INFORMS).

[19] Defazio A, Bach F, Lacoste-Julien S (2013) SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. Ghahramani Z, Welling M, Cortes C, Lawrence N, Weinberger K, eds., *Advances in Neural Information Processing Systems 27*, 1646–1654 (Curran Associates, Inc.).

[20] Delyon B (2009) Exponential inequalities for sums of weakly dependent variables. *Electronic journal of probability* 14(none):752–779, ISSN 1083-6489.

[21] Deng G, Ferris MC (2009) Variable-number sample-path optimization. *Mathematical Programming* (117):81–109.

[22] Friedlander MP, Schmidt M (2012) Hybrid deterministic-stochastic methods for data fitting. *SIAM Journal on Scientific Computing* 34(3):A1380–A1405, URL http://dx.doi.org/10.1137/110830629.

[23] Fu MC, Glover FW, April J (2005) Simulation optimization: A review, new developments, and applications. *Proceedings of the Winter Simulation Conference* (IEEE), URL http://dx.doi.org/10.1109/wsc.2005.1574242.

[24] Ghadimi S, Lan G (2013) Stochastic First- and Zeroth-Order Methods for Nonconvex Stochastic Programming. *SIAM J. Optim.* 23(4):2341–2368, ISSN 1052-6234, URL http://dx.doi.org/10.1137/120880811.

[25] Ghanem RG, Spanos PD (1991) *Stochastic Finite Elements: A Spectral Approach* (New York, NY: Springer New York), ISBN 9781461277958.

[26] Giné E, Nickl R (2016) *Mathematical foundations of infinite-dimensional statistical models*. Cambridge series on statistical and probabilistic mathematics, ISBN 9781107043169.

[27] Hashemi FS, Ghosh S, Pasupathy R (2014) On adaptive sampling rules for stochastic recursions. *Proceedings of the Winter Simulation Conference 2014*, 3959–3970, URL http://dx.doi.org/10.1109/WSC.2014.7020221.

[28] Healy K, Schruben LW (1991) Retrospective simulation response optimization. Nelson BL, Kelton DW, Clark GM, eds., *Proceedings of the 1991 Winter Simulation Conference*, 954–957 (Institute of Electrical and Electronics Engineers: Piscataway, New Jersey).

[29] Homem-de-Mello T (2003) Variable-sample methods for stochastic optimization. *ACM Transactions on Modeling and Computer Simulation* 13:108–133.

[30] Huber PJ (1992) Robust estimation of a location parameter. *Breakthroughs in statistics*, 492–518 (Springer).

[31] Jalilzadeh A, Nedić A, Shanbhag UV, Yousefian F (2021) A variable sample-size stochastic quasi-newton method for smooth and nonsmooth stochastic convex optimization. *Mathematics of Operations Research* .

[32] Jalilzadeh A, Shanbhag UV (2016) eg-vssa: An extragradient variable sample-size stochastic approximation scheme: Error analysis and complexity trade-offs. *2016 Winter Simulation Conference (WSC)*, 690–701 (IEEE).

[33] Johnson R, Zhang T (2013) Accelerating stochastic gradient descent using predictive variance reduction. *Advances in neural information processing systems* 26.

[34] Juditsky A, Lan G, Nemirovski A, Shapiro A (2009) Stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization* 19(4):1574–1609.

[35] Kim S, Pasupathy R, Henderson SG (2014) A guide to SAA. Fu M, ed., *Encyclopedia of Operations Research and Management Science*, Hillier and Lieberman OR Series (Elsevier).

[36] Kim S, Pasupathy R, Henderson SG (2015) A guide to sample average approximation. Fu M, ed., *Handbook of Simulation Optimization*, volume 216 of *International Series in Operations Research & Management Science*, 207–243 (Springer), URL http://dx.doi.org/10.1007/978-1-4939-1384-8_8.

[37] Kingma DP, Ba J (2014) ADAM: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

[38] Kreyszig E (1989) *Introductory functional analysis with applications*. Wiley classics library (New York: Wiley), wiley classics library ed. edition, ISBN 0471504599.

[39] Kushner HJ, Yin GG (2003) *Stochastic Approximation and Recursive Algorithms and Applications* (New York, NY: Springer-Verlag).

[40] Lan G (2011) An optimal method for stochastic composite optimization. *Mathematical programming* 133(1-2):365–397, ISSN 0025-5610.

[41] Lydia A, Francis S (2019) Adagrad—an optimizer for stochastic gradient descent. *Int. J. Inf. Comput. Sci* 6(5):566–568.

[42] Mak WK, Morton DP, Wood RK (1999) Monte Carlo bounding techniques for determining solution quality in stochastic programs. *Operations Research Letters* 24:47–56.

[43] Mania H, Guy A, Recht B (2018) Simple random search of static linear policies is competitive for reinforcement learning. Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R, eds., *Advances in Neural Information Processing Systems 31*, 1800–1809 (Curran Associates, Inc.), URL http://papers.nips.cc/paper/7451-simple-random-search-of-static-linear-policies-is-competitive-for-reinforcement-learning.pdf.

[44] Mei S, Bai Y, Montanari A (2018) The landscape of empirical risk for nonconvex losses. *The Annals of statistics* 46(6A):2747–2774, ISSN 0090-5364.

[45] Mokhtari A, Ozdaglar A, Jadbabaie A (2019) Efficient nonconvex empirical risk minimization via adaptive sample size methods. *The 22nd International Conference on Artificial Intelligence and Statistics*, 2485–2494 (PMLR).

[46] Mokhtari A, Ribeiro A (2015) Global convergence of online limited memory bfgs. *Journal of Machine Learning Research* 16(1):3151–3181.

[47] Nelson BL (2013) *Foundations and Methods of Stochastic Simulation: A First Course* (New York, NY: Springer).

[48] Nemirovskii A, Juditsky A, Lan G, Shapiro A (2009) Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization* 19(4):1574–1609.

[49] Nemirovsky AS, Yudin DB (1983) *Problem complexity and method efficiency in optimization* (New York, NY: Wiley).

[50] Nesterov Y (2004) *Introductory Lectures on Convex Optimization: A Basic Course* (New York, NY: Springer Science + Business Media, LLC).

[51] Nocedal J, Wright S (1999) *Numerical Optimization* (Springer New York), 2 edition.

[52] Pasupathy R (2010) On choosing parameters in retrospective-approximation algorithms for stochastic root finding and simulation optimization. *Operations Research* 58:889–901.

[53] Pasupathy R, Ghosh S (2013) Simulation optimization: A concise overview and implementation guide. *Theory Driven by Influential Applications*, 122–150 (INFORMS), URL http://dx.doi.org/10.1287/educ.2013.0118.

[54] Pasupathy R, Glynn P, Ghosh S, Hashemi FS (2018) On sampling rates in simulation-based recursions. *SIAM Journal on Optimization* 28(1):45–73, URL http://dx.doi.org/10.1137/140951679.

[55] Pasupathy R, Kim S (2011) The stochastic root-finding problem: Overview, solutions, and open questions. *ACM Transactions on Modeling and Computer Simulation* 21(3):1–23.

[56] Pasupathy R, Song Y (2021) An adaptive sequential sample average approximation framework for solving two-stage stochastic programs. *SIAM Journal on Optimization* 31(1):1017–1048.

[57] Polyak BT, Juditsky AB (1992) Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization* 30(4):838–855.

[58] Robbins H, Monro S (1951) A stochastic approximation method. *Annals of Mathematical Statistics* 22:400–407.

[59] Roosta-Khorasani F, Mahoney MW (2019) Sub-sampled Newton methods. *Mathematical Programming* 174(1-2):293–326, URL http://dx.doi.org/10.1007/s10107-018-1346-5.

[60] Rubinstein RY, Shapiro A (1990) Optimization of static simulation models by the score function method. *Mathematics and Computers in Simulation* 32:373–392.

[61] Rustagi JS (2014) *Optimization techniques in statistics* (Elsevier).

[62] Salimans T, Ho J, Chen X, Sidor S, Sutskever I (2017) Evolution strategies as a scalable alternative to reinforcement learning. Technical Report 1703.03864, ArXiv, URL https://arxiv.org/abs/1703.03864.

[63] Schmidt M, Le Roux N, Bach F (2017) Minimizing finite sums with the stochastic average gradient. *Mathematical Programming* 162(1):83–112.

[64] Schraudolph NN, Yu J, Günter S (2007) A stochastic quasi-newton method for online convex optimization. *International Conference on Artificial Intelligence and Statistics*, 436–443.

[65] Shanbhag UV, Blanchet JH (2015) Budget-constrained stochastic approximation. *2015 Winter Simulation Conference (WSC)*, 368–379 (IEEE).

[66] Shapiro A (1991) Asymptotic analysis of stochastic programs. *Annals of Operations Research* 30:169–186.

[67] Shapiro A (1993) Asymptotic behavior of optimal solutions in stochastic programming. *Mathematics of Operations Research* 18(4):829–845.

[68] Shapiro A (2000) Statistical inference of stochastic optimization problems. Uryasev SP, ed., *Probabilistic Constrained Optimization: Methodology and Applications*, 282–304 (Kluwer Academic Publishers).

[69] Shapiro A (2004) Monte Carlo sampling methods. Ruszczynski A, Shapiro, eds., *Stochastic Programming*, 353–426, Handbooks in Operations Research and Management Science (Elsevier).

[70] Shapiro A, Dentcheva D, Ruszczynski A (2009) *Lectures on Stochastic Programming: Modeling and Theory* (Philadelphia, Pennsylvania: Society for Industrial and Applied Mathematics).

[71] Shapiro A, Kleywegt A (2001) Minimax analysis of stochastic problems. *Optimization Methods and Software* 17:523–542.

[72] Stefanski LA, Boos DD (2002) The calculus of m-estimation. *The American Statistician* 56(1):29–38.

[73] Talagrand M (1996) New concentration inequalities in product spaces. *Inventiones mathematicae* 126(3):505–563.

[74] Trosset M (2009) *An Introduction to Statistical Inference and Its Applications with R* (New Jersey: CRC/Chapman & Hall).

[75] van de Geer SA (2006) *Empirical Processes in M-Estimation*. Cambridge Series in Statistical and Probabilistic Mathematics, Series Number 6 (Cambridge: Cambridge University Press), 1-st edition, ISBN 0521123259.

[76] van der Vaart AW, Wellner JA (1996) *Weak Convergence and Empirical Processes: With Applications to Statistics* (New York, New York: Springer-Verlag).

[77] Vapnik VN (1995) *The nature of statistical learning theory* (New York: Springer), ISBN 1-4757-2440-3.

[78] Vershynin R (2018) *High-Dimensional Probability: An Introduction with Applications in Data Science* (Cambridge Series in Statistical and Probabilistic Mathematics).

[79] Ward R, Wu X, Bottou L (2018) Adagrad stepsizes: Sharp convergence over nonconvex landscapes .

[80] Williams D (1991) *Probability with Martingales* (Cambridge, UK: Cambridge University Press).

SUPPLEMENTARY MATERIAL

Appendix A: Alternate Iteration Complexity   In this section, we provide the proof of Theorem 4 and also an example to illustrate its applicability.

Proof of Theorem 4. Recall that

$$R_{M_k}(X_k) := \frac{\nabla f_{M_k}(X_k) - \nabla f(X_k)}{\|\nabla f(X_k)\| + c_0}. \tag{109}$$

Taking norm and applying triangle inequality to (109), we can write

$$\|\nabla f(X_k)\| \le \|\nabla f_{M_k}(X_k)\| + c_0\|R_{M_k}(X_k)\| + \|R_{M_k}(X_k)\|\|\nabla f(X_k)\|. \tag{110}$$

Take expectation on both sides of (110) and noticing that $\|\nabla f_{M_k}(X_k)\| \le \epsilon_k$, we can write

$$\mathbb{E}\left[\|\nabla f(X_k)\|\right] \le \epsilon_k + c_0\mathbb{E}\left[\|R_{M_k}(X_k)\|\frac{\|\nabla f(X_k)\|^{\alpha} + \delta_0}{\|\nabla f(X_k)\|^{\alpha} + \delta_0}\right] + \mathbb{E}\left[\|R_{M_k}(X_k)\|\|\nabla f(X_k)\|\frac{\|\nabla f(X_k)\|^{\alpha} + \delta_0}{\|\nabla f(X_k)\|^{\alpha} + \delta_0}\right]$$

$$\le \epsilon_k + \frac{c_0}{\delta_0}\mathbb{E}\left[\|R_{M_k}(X_k)\|(\|\nabla f(X_k)\|^{\alpha} + \delta_0)\right]$$

$$+ \mathbb{E}\left[\underbrace{\|R_{M_k}(X_k)\|(\|\nabla f(X_k)\|^{\alpha} + \delta_0)}_{A}\underbrace{\|\nabla f(X_k)\|^{1-\alpha}}_{B}\right]. \tag{111}$$

Recall from Young's inequality [38, pp. 13] that

$$AB \le \frac{\gamma^p A^p}{p} + \frac{B^q}{\gamma^q q}; \quad \frac{1}{p} + \frac{1}{q} = 1; \quad p, q \ge 1; \quad \gamma > 0. \tag{112}$$

Apply (112) to the product of the quantities designated A and B in (111), choose

$$p := \frac{1}{\alpha}; \quad q := \frac{1}{1-\alpha}; \quad \gamma := 2(1-\alpha)^{1-\alpha},$$

and take expectation to get

$$\mathbb{E}\left[\|R_{M_k}(X_k)\|(\|\nabla f(X_k)\|^{\alpha} + \delta_0)\|\nabla f(X_k)\|^{1-\alpha}\right]$$

$$\le \frac{\alpha(2(1-\alpha))^{\frac{1}{\alpha}}}{1-\alpha}\mathbb{E}\left[\left(\|R_{M_k}(X_k)\|(\|\nabla f(X_k)\|^{\alpha} + \delta_0)\right)^{1/\alpha}\right] + \frac{1}{2}\mathbb{E}\left[\|\nabla f(X_k)\|\right]. \tag{113}$$

Plugging (113) in (111), we get

$$\frac{1}{2}\mathbb{E}\left[\|\nabla f(X_k)\|\right] \le \epsilon_k + \frac{c_0}{\delta_0}\mathbb{E}\left[\|R_{M_k}(X_k)\|(\|\nabla f(X_k)\|^{\alpha} + \delta_0)\right]$$

$$+ \frac{\alpha(2(1-\alpha))^{\frac{1}{\alpha}}}{1-\alpha}\mathbb{E}\left[\left(\|R_{M_k}(X_k)\|(\|\nabla f(X_k)\|^{\alpha} + \delta_0)\right)^{1/\alpha}\right]$$

$$\le \epsilon_k + \frac{c_0}{\delta_0}\mathbb{E}\left[\mathbb{E}\left[\sup_{x \in \mathbb{R}^d}\|R_{M_k}(x)\|(\|\nabla f(x)\|^{\alpha} + \delta_0)\mid \mathscr{F}_{k-1}\right]\right]$$

$$+ \frac{\alpha(2(1-\alpha))^{\frac{1}{\alpha}}}{1-\alpha}\mathbb{E}\left[\mathbb{E}\left[\left(\sup_{x \in \mathbb{R}^d}\|R_{M_k}(x)\|(\|\nabla f(x)\|^{\alpha} + \delta_0)\right)^{1/\alpha}\mid \mathscr{F}_{k-1}\right]\right]$$

$$= \epsilon_k + \frac{c_0}{\delta_0}\mathbb{E}\left[\mathbb{E}\left[H_{M_k}\mid \mathscr{F}_{k-1}\right]\right] + \frac{\alpha(2(1-\alpha))^{\frac{1}{\alpha}}}{1-\alpha}\mathbb{E}\left[\mathbb{E}\left[H_{M_k}^{1/\alpha}\mid \mathscr{F}_{k-1}\right]\right]. \tag{114}$$

Due to the assumptionin (40), and since $\alpha \in (0,1)$, we know that there exists $\sigma_0 < \infty$ such that

$$\mathbb{E}\left[\left(\sqrt{M_k}\,H_{M_k}\right)^n \mid \mathscr{F}_{k-1}\right] \leq \sigma_0^n \text{ for } n \leq \frac{1}{\alpha}. \tag{115}$$

Plugging (115) in (114) and since $\epsilon_k := C_{2,k}/M_k^p$ where $\mathscr{F}_{k-1} \ni C_{2,k} \leq c_2$, we get

$$\mathbb{E}\left[\|\nabla f(X_k)\| \mid \mathscr{F}_{k-1}\right] \leq \frac{2c_2}{M_k^p} + \frac{2c_0\sigma_0/\delta_0}{\sqrt{M_k}} + \left(\frac{\kappa_2}{\sqrt{M_k}}\right)^{1/\alpha}, \tag{116}$$

proving (41). Now plug the choice of sample size sequence given in (C.1(a)) in (116) and $p = 1/2$ to see that the iteration complexity (36) holds as well.

The key assumption of the iteration complexity result in Theorem 4 is that the envelope of the weighted relative error class $w\mathscr{R} := \{R(x)(\|\nabla f(x)\|^\alpha + \delta_0), x \in \mathscr{X}\}$ exhibits the CLT-scaling (40). How stringent is this assumption? In particular, how does it relate to the classes introduced through the various example classes discussed thus far? The following example answers this question from the standpoint of the simple but important Example 6.2 since it provides the most intuition.

Example (Power Error Class). Suppose the random error function $\nabla F(\cdot, Y) - \nabla f(\cdot)$ is such that for some $\gamma_0 \in (0,1)$,

$$\nabla F(x,Y) - \nabla f(x) = A(Y)\|\nabla f(x)\|^{1-\gamma_0} + B(Y), \quad x \in \mathbb{R}^d \tag{117}$$

where $A(Y)$ and $B(Y)$ are mean zero sub-exponential random variables.

Let's now demonstrate that the class characterized by (117) (along with sub-exponential $A(Y), B(Y)$ assumption) satisfies the postulates of Theorem 4, that is, its weighted relative error class has an empirical-mean envelope satisfying (40). Choosing $\alpha < \gamma_0$, and denoting $\bar{A}_m = m^{-1}\sum_{j=1}^m A(Y_j)$, $\bar{B}_m = m^{-1}\sum_{j=1}^m B(Y_j)$ we see that

$$\|R_m(x)(\|\nabla f(x)\|^\alpha + \delta_0)\| \leq \left\|\frac{\bar{A}_m\|\nabla f(x)\|^{1-\gamma_0}}{\|\nabla f(x)\| + c_0}(\|\nabla f(x)\|^\alpha + \delta_0)\right\| + \left\|\frac{\bar{B}_m}{\|\nabla f(x)\| + c_0}(\|\nabla f(x)\|^\alpha + \delta_0)\right\|$$
$$\leq (1+\delta_0)(1 \vee \frac{1}{c_0})(|\bar{A}_m| + |\bar{B}_m|). \tag{118}$$

We see from (118) that the empirical-mean envelope $H_m := \|R_m(x)(\|\nabla f(x)\|^\alpha + \delta_0)\|$ satisfies

$$\mathbb{E}[m^{1/2\alpha}H_m^{1/\alpha}] \leq \left((1+\delta_0)(1 \vee \frac{1}{c_0})\right)^{1/\alpha} 2^{1/\alpha} m^{1/2\alpha}\left(\mathbb{E}\left[|\bar{A}_m|^{1/\alpha}\right] + \mathbb{E}\left[|\bar{B}_m|^{1/\alpha}\right]\right)$$
$$\leq \left(2(1+\delta_0)(1 \vee \frac{1}{c_0})\right)^{1/\alpha} m^{1/2\alpha}\left(\mathbb{E}\left[|\bar{A}_m|^{1/\alpha}\right] + \mathbb{E}\left[|\bar{B}_m|^{1/\alpha}\right]\right). \tag{119}$$

Since $A(Y)$ is a sub-exponential random variable, and since $\bar{A}_m$ is the sample mean of iid random variables, there exist constants $c_A, \sigma_A \in (0, \infty)$ such that

$$\mathbb{E}\left[|\bar{A}_m|^{1/\alpha}\right] = \int_0^\infty P\left(\left|\sum_{j=1}^m A(Y_j)\right| > mt^\alpha\right)$$
$$\leq \int_0^{\sigma_A^2 c_A} 2\exp\{-\frac{1}{2}\frac{m}{\sigma_A^2}t^{2\alpha}\}\,dt + \int_{\sigma_A^2 c_A}^\infty 2\exp\{-\frac{1}{2}mc_At^\alpha\}\,dt$$
$$\leq \frac{1}{m^{1/2\alpha}}\left(\kappa_{1A} + \frac{\kappa_{2A}}{m^{1/2\alpha}}\right), \tag{120}$$

where

$$\kappa_{1A} := \frac{2}{\alpha}\left(\frac{2}{c_A}\right)^{1/\alpha}\Gamma(\frac{1}{\alpha}); \quad \kappa_{2A} := \frac{1}{\alpha}\left(\frac{2\sigma_A^2}{c_A}\right)^{1/2\alpha}\Gamma(\frac{1}{2\alpha}),$$

and the second inequality in (120) follows the standard tail bound on sub-exponential random variables [78]. Similarly, since $B(Y)$ is a sub-exponential random variable, and since $\bar{B}_m$ is the sample mean of iid random variables, there exist constants $c_B, \sigma_B \in (0, \infty)$ such that

$$\mathbb{E}\left[\left|\bar{B}_m\right|^{1/\alpha}\right] \leq \frac{1}{m^{1/2\alpha}}\left(\kappa_{1B} + \frac{\kappa_{2B}}{m^{1/2\alpha}}\right), \tag{121}$$

where

$$\kappa_{1B} := \frac{2}{\alpha}\left(\frac{2}{c_B}\right)^{1/\alpha}\Gamma(\frac{1}{\alpha}); \quad \kappa_{2B} := \frac{1}{\alpha}\left(\frac{2\sigma_B^2}{c_B}\right)^{1/2\alpha}\Gamma(\frac{1}{2\alpha}).$$

Plugging (120) and (121) in (119), we see that $\mathbb{E}[m^{1/2\alpha}H_m^{1/\alpha}]$ is finite if $\alpha < \gamma_0$ and we conclude that the power error class considered in this example satisfies the postulates of Theorem 4. ∎

Appendix B: Key Results from the Theory of Empirical Processes   In this section, we provide some relevant results that relate the (ULLN), or (R-ULLN) to the underlying geometric or regularity property of the functions generated by the oracle called upon in our algorithm. Note that the crucial property is the concept of "uniform" in the Law of Large Numbers within a class of functions. Such a theory has been developed quite extensively in the realm of empirical processes. We refer to [26, 75, 76] for comprehensive treatments.

Recall that by classical Law of Large Numbers, given any probability measure P and function F with $P|F| < \infty$, we have that

$$P_n F := \frac{1}{n}\sum_{i=1}^{n}F(Y_i) \longrightarrow_{n\to\infty} PF := \mathbb{E}F(Y), \quad P \text{ a.s.}$$

where Y_i are IID samples with distribution given by P. The notation P_n refers to the empirical measure with respect to the samples. We would like the above limit to hold uniformly for a class $\mathscr{F}$ of functions. This is captured by the following uniform norm:

$$\|P_n - P\|_{\mathscr{F}} := \sup_{F \in \mathscr{F}}|P_n F - PF|$$

Then the Uniform Law of Large Numbers (ULLN) with respect to the class $\mathscr{F}$ is stated as:

$$\lim_{n\to 0}\|P_n - P\|_{\mathscr{F}} = 0, \quad P \text{ a.s.}$$

If the above is true, then $\mathscr{F}$ is called a *Glivenko-Cantelli* class.

Translating the above framework to our current RA setting, F and $\mathscr{F}$ are specifically given by

$$\mathscr{F} = \left\{F_x := F(x, \cdot) : x \in \mathscr{X} \ (\text{or } \mathbb{R}^d)\right\}$$

so that the functions are indexed by $x \in \mathscr{X}$. We will further consider the relative error class (9) repeated here for convenience:

$$\mathscr{R} := \left\{R_x := R(x, \cdot), \quad x \in \mathscr{X} \ (\text{or } \mathbb{R}^d)\right\} \quad \text{where} \quad R(x, Y) := \frac{\nabla F(x, Y) - \nabla f(x)}{\|\nabla f(x)\| + c_0}.$$

In the above cases, we have

$$\begin{cases} P_n F_x & = \frac{1}{n}\sum_{i=1}^{n}F(x, Y_i), \\ PF_x & = \mathbb{E}F(x, Y) \ (:= f(x)), \\ \|P_n - P\|_{\mathscr{F}} & = \sup_{x \in \mathscr{X}}|P_n F_x - f(x)| \end{cases} \quad \text{and} \quad \begin{cases} P_n R_x & = \frac{\frac{1}{n}\sum_{i=1}^{n}F(x, Y_i) - \nabla f(x)}{\|\nabla f(x)\| + c_0}, \\ PR_x & = \mathbb{E}R(x, Y) \ (= 0), \\ \|P_n - P\|_{\mathscr{R}} & = \sup_{x \in \mathscr{X}}|P_n R_x| \end{cases}$$

Hence, the Uniform Law of Large Numbers is exactly given by

$$\lim_{n\to\infty} \sup_{x\in\mathscr{X}} \left| \frac{1}{n}\sum_{i=1}^{n}\nabla F(x,Y_i) - \nabla f(x) \right| = 0, \quad \text{and} \quad \lim_{n\to\infty} \sup_{x\in\mathscr{X}} \left| \frac{\frac{1}{n}\sum_{i=1}^{n}\nabla F(x,Y_i) - \nabla f(x)}{\|\nabla f(x)\| + c_0} \right| = 0.$$

But in order to keep this Appendix in a general setting, we will follow the usual notation in empirical processes, using F and $\mathscr{F}$ to denote a particular function and the function class.

As mentioned before, the essential criterior for the validity of ULLN is given by the entropy of $\mathscr{F}$ which measures the "(ir)regularity" or "richness" of the function class. This is given by the following definition.

DEFINITION 6 (EMPIRICAL ENTROPY).   A δ-net for $\mathscr{F}$ is a set $\{F_1, F_2, \ldots, F_n\}$ of functions in $\mathscr{F}$ such that any F can be approximated to within δ in $L_p(P)$ norm ($p \geq 1$) using one of the F_j's. That is, for any $F \in \mathscr{F}$, there is some $j \in \{1, 2, \ldots, n\}$ such that,

$$\left(\mathbb{E}\left[\|F(Y) - F_j(Y)\|^p \right] \right)^{1/p} = \left(\int \|F(Y) - F_j(Y)\|^p \, \mathrm{d}P \right)^{1/p} \leq \delta.$$

The δ-*covering number* $N_p(\delta, \mathscr{F}, P)$ of the class $\mathscr{F}$ is then the "smallest" δ-net, i.e.,

$$N_p(\delta, \mathscr{F}, P) := \inf\left\{ n : \text{there exists a } \delta\text{-net of } \mathscr{F} \text{ with cardinality } n \right\}.$$

The *entropy* of the class $\mathscr{F}$ is the logarithm of its δ-covering number:

$$H_p(\delta, \mathscr{F}, P) = \log N_p(\delta, \mathscr{F}, P).$$

In actual application, the empirical version of the above is often used. More precisely, the measure P is replaced by its empirical measure P_n, leading to the following *empirical entropy*:

$$H_p(\delta, \mathscr{F}, P_n).$$

Note that the above is a *random* quantity. We further recall the definition here of the envelope G given in (11):

$$G = G(Y) := \sup\left\{ \|F(Y)\| : F \in \mathscr{F} \right\}. \tag{11}$$

It is interesting to note that by just assuming $\mathbb{E}G < \infty$, we already have that $\|P_n - P\|_{\mathscr{R}}$ converges to some *finite limit* P a.s. and in L^1 [26, Prop. 3.7.8].

To show that the limit is actually zero, a growth restriction on H_p is crucial. More precisely, let

$$\mathscr{F}_B = \{F\mathbb{I}_{F\leq B} : F \in \mathscr{F}\}, \quad \text{where } F\mathbb{I}_{F\leq B} = \begin{cases} F(x,Y), & \|F(x,Y)\| \leq M, \\ 0, & \|F(x,Y)\| > M. \end{cases}$$

Then by [26, Thm. 3.7.14], under the assumption that $\mathscr{F}$ is L^1-bounded, i.e. $\sup_{F\in\mathscr{F}} \mathbb{E}|F| < \infty$, the fact that $\mathscr{F}$ being Glivenko-Cantelli is equivalent to the following statement:

$\mathbb{E}G < \infty$, and for all $B < \infty$ and $\delta > 0$, and some $p \in (0, \infty]$, it holds that

$$\frac{1}{n} H_p(\delta, \mathscr{F}, P_n) \xrightarrow{\text{P}} 0 \text{ as } n \to \infty. \tag{Sub-Ent}$$

By [26, Corollary 3.7.15], the above is also equivalent to

$$\mathbb{E}G < \infty \text{ and } \frac{1}{n}\mathbb{E}\sqrt{H_2(\delta, \mathscr{F}, P_n)} \longrightarrow 0. \tag{Sub-Ent$_2$}$$

See also [76, Theorem 2.4.3] for a similar statement.

We make the following remarks pertaining to the analysis of RA.

(a) An implication of such equivalence is that the main condition (R-ULLN) in Theorem 1 can be replaced by (Sub-Ent) without changing the assertions of the theorem. Even though (Sub-Ent) is one of the key "workhorse" conditions in empirical process theory, we have chosen to present our results with (R-ULLN) since, as we shall demonstrate in Section 6, the condition in (R-ULLN) appears to be much easier to check in optimization contexts.

(b) Following the proofs of [26, Thm. 3.7.14, Corollary 3.7.15] and [76, Theorem 2.4.3], we can extend the results to second or higher order moments and with respect to a filtration. In particular, using the notations of Theorems 1 and 2, under the assumption $\mathbb{E}[G] < \infty$, the sublinear entropy growth condition (Sub-Ent) can be used to show that

$$\sup_x \|R_{M_k}(x)\| \overset{\text{wp1}}{\to} 0 \quad \text{and} \quad \mathbb{E}\left[\sup_x \|R_{M_k}(x)\| \,|\, \mathscr{F}_{k-1}\right] \to 0 \quad \text{if} \quad M_k \to \infty.$$

Furthermore, if $\mathbb{E}[G^2] < \infty$, then we can have

$$\mathbb{E}\left[\sup_x \|R_{M_k}(x)\|^2 \,|\, \mathscr{F}_{k-1}\right] \to 0 \quad \text{if} \quad M_k \to \infty.$$

This is something we will use in Theorem 2 when we demonstrate the L_1-consistency.

The next question of theoretical and practical importance is the *rate* of convergence in (ULLN). By the classical Central Limit Theorem (CLT), we expect $\|P_n - P\|_{\mathscr{F}} \lesssim O(\frac{1}{\sqrt{n}})$. Again, the uniformity in $\mathscr{F}$ is crucial in our application. This is achieved by means of the following quantity [26, p. 186].

DEFINITION 7 (KOLTCHINSKII-POLLARD ENTROPY). Given a class of function $\mathscr{R}$ and its envelope function G,

$$J(\mathscr{F}, G, \delta) = \int_0^\delta \sup_Q \sqrt{H_2\left(\tau \|G\|_{L^2(Q)}, \mathscr{F}, Q\right)} \, d\tau \tag{122}$$

where the sup is taken over all finite discrete probabilities Q with $\|G\|_{L^2(Q)} > 0$.

With the above, we have the following theorem which exactly gives the type of CLT scaling we need [26, Theorem 3.5.4].

THEOREM 8. *The following estimate holds for all $n \in \mathbb{N}$:*

$$\mathbb{E}\left[\sqrt{n} \|P_n - P\|_{\mathscr{F}}\right] \leq \max\left\{A_1 \|G\|_{L^2(P)} J(\mathscr{F}, G, \delta), \frac{A_2 \|U\|_{L^2(P)} J^2(\mathscr{F}, G, \delta)}{\sqrt{n}\delta^2}\right\} \tag{123}$$

where

$$U = \max_{1 \leq i \leq n} G(Y_i), \quad \sigma^2 = \sup_{F \in \mathscr{F}} PF^2, \quad \delta = \frac{\sigma}{\|G\|_{L^2(P)}},$$

and A_1 and A_2 are some explicit universal constants.

REMARK 7. Note that all of the above results about the limiting behavior of empirical measures make use of the entropy function H_p. Readers might be aware of another concept which can provide explicit bounds on H_p. It is given by means of VC dimension, named after Vladimir Vapnik and Alexey Chervonenkis. VC dimension measures the combinatorial richness of the class of functions in question. Without getting too much into detail, we will simply state the following definition which illustrates the type of bounds we can have on H_p.

DEFINITION 8. [26, Definition 3.6.10] A class of measurable functions $\mathscr{F}$ is of *VC* type with respect to a envelope function G if there exist finite constants A and ν such that for all probability meaures Q and $0 < \epsilon < 1$, it holds that

$$N_2(\epsilon \|G\|_{L^2(Q)}, \mathscr{F}, Q) \leq \left(\frac{A}{\epsilon}\right)^\nu.$$

The above estimate clealy provides sufficient bound for the entropy function that will be workable for all our results.

The following concentration inequality by Talagrand plays a fundamental role in the theory of empirical processes.

THEOREM 9.   *[73, Theorem 1.4] Using the previous notations, in particular, let $Y_i, \ldots, Y_n$ be IID variables distributed according to P. Let $\mathscr{F}$ be any class of functions that is uniformly bounded by a constant $U > 0$. Then for all $t > 0$,*

$$P\left\{\left|\left\|\sum_{i=1}^{n} F(Y_i)\right\|_{\mathscr{R}} - \mathbb{E}\left\|\sum_{i=1}^{n} F(Y_i)\right\|_{\mathscr{R}}\right| \geq t\right\} \leq K \exp\left\{-\frac{1}{K}\frac{t}{U}\log\left(1 + \frac{tU}{V}\right)\right\} \qquad (124)$$

where K is a universal constant and V is any number satisfying

$$V \geq \mathbb{E}\sup_{F \in \mathscr{F}}\sum_{i=1}^{n} F^2(X_i)$$

In order to see the explicit scaling on n, we can use $V = nU^2$. Upon replacing t by $\frac{t}{n}$, then (124) becomes

$$P\left\{\left|\left\|\frac{1}{n}\sum_{i=1}^{n} F(Y_i)\right\|_{\mathscr{F}} - \mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n} F(Y_i)\right\|_{\mathscr{F}}\right| \geq t\right\} \leq K \exp\left\{-\frac{1}{K}\frac{nt}{U}\log\left(1 + \frac{t}{U}\right)\right\}. \qquad (125)$$

This is the form used in the proof of Theorem 2.